Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.081705

ARTICLE

Hybrid Ensemble and Federated Learning Framework for Privacy-Preserving Cardiovascular MRI Segmentation

Karim Gasmi[1,*], Afrah Alanazi[2], Inam Alanazi[2], Sahar Almenwer[1], Norah Alanazi[1], Sarah Almaghrabi[3] and Samia Yahyaoui[4]

[1]Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia
[2]Department of Information System, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia
[3]Department of Information Technology, College of Computing and Information Technology at Khulais, University of Jeddah, Jeddah, Saudi Arabia
[4]Department of Physics, College of Science, Jouf University, Sakaka, Saudi Arabia

*Corresponding Author: Karim Gasmi. Email: kgasmi@ju.edu.sa

Received: 07 March 2026; Accepted: 19 May 2026; Published: 30 June 2026

ABSTRACT: Cardiac magnetic resonance imaging (MRI) segmentation is an essential aspect of quantitative cardiovascular analysis, facilitating accurate evaluation of ventricular volumes, myocardial mass, and functional parameters. Deep learning-based segmentation models have shown strong performance on benchmark datasets such as ACDC, but they remain challenging to deploy in real-world multi-centre settings. Data privacy laws make it hard to share data across institutions, and differences in imaging protocols and patient populations mean that data is not always distributed in the same way (non-IID). This can have a big impact on how well models work together and how well they generalise. To address these issues, we first evaluate advanced segmentation architectures, including UNet++ and FPN with EfficientNet-based encoders, and assess multiple hybrid combinations at the probability level. We further improve the ensemble strategy by using a genetic algorithm to automatically identify the optimal model-weighting scheme, rather than fixed combination coefficients. The genetic algorithm explores the solution space to identify the optimal weight configuration based on segmentation metrics. The best hybrid configuration is then chosen as the input architecture for the federated learning stage. We propose a privacy-preserving federated ensemble framework that enables multiple clients to collaboratively train segmentation models without sharing raw MRI data. We methodically evaluate three federated optimisation strategies: FedAvg under IID and non-IID client distributions, and FedProx, which incorporates proximal regularisation to reduce client drift. The genetically optimised ensemble is always used in all federated setups. A thorough analysis of ACDC testing volumes employing overlap- and boundary-based metrics illustrates that the amalgamation of hybrid learning with genetic optimisation and federated training enhances robustness in heterogeneous environments while maintaining data confidentiality, thus providing an efficient approach for secure multi-centre cardiac MRI segmentation.

KEYWORDS: SDG 3; cardiovascular imaging; segmentation; ensemble deep learning; genetic algorithm; federated learning; privacy-preserving AI

1 Introduction

Cardiovascular disease (CVD) is still the number one cause of death in the world. According to the World Health Organisation, 19.8 million people died from CVD in 2022, which is about 32% of all deaths worldwide. About 85% of these deaths were caused by heart attack and stroke [1]. The World

Heart Federation also says that 20.5 million people died from CVD in 2021, which is almost one-third of all deaths in the world [2]. These numbers demonstrate that CVD continues to harm people and the economy, underscoring the importance of imaging pipelines that can produce reliable, reproducible biomarkers at scale. In cardiac MR (CMR), CT angiography (CTA), X-ray coronary angiography (CAG), and echocardiography, segmentation is the first step that converts pixel data into clinical measurements, such as chamber volumes, ejection fraction (EF), myocardial thickness, aortic diameters, or coronary lumen geometry. At this point, any bias will affect the derived indices and subsequent decisions. On the other hand, strong segmentation supports consistent diagnosis, risk stratification, and therapy planning [3].

Recent developments in artificial intelligence (AI) have transformed cardiovascular image analysis. AI techniques reduce reporting time, reduce differences between and within observers, and leverage multi-slice, multi-phase, or multi-view context to stabilise challenging areas (e.g., apical/basal CMR slices, valve planes, and small-calibre coronary arteries). The community has reached consensus on a set of reporting standards to compare segmentation quality across modalities and groups. These standards include Dice/IoU for overlap, Hausdorff/ASSD for boundary, and clinical agreement (e.g., EF error, diameter deviation) [4,5].

Historically, machine learning (ML) and model-based approaches constituted the foundation of cardiac segmentation: pixel classification, deformable and level-set models incorporating region and gradient cues, shortest-path formulations, and robust priors through statistical shapes or atlas registration [3]. These pipelines are easy to understand and use few labels, but they often need careful setup and manual tuning. They can also break easily if the vendor or protocol changes, especially in low-contrast or highly remodelled anatomy. In contrast, deep learning (DL) redefined segmentation as end-to-end representation learning utilising U-Net-like encoder-decoder architectures, multi-scale attention mechanisms, and temporal or multi-view fusion. DL systems consistently enhance overlap and boundary metrics across CMR, CTA, echo, and CAG, establishing themselves as the de facto standard for multi-structure cardiac segmentation [4,6,7].

First, privacy and governance rules make it difficult to share data across multiple centres, hindering generalisation and slowing validation in representative populations. Second, a change in domain (scanner, protocol, contrast phase, population) can hurt performance outside of the development cohort. Reviews stress the need for standard practices like harmonisation and domain adaptation [4,5]. Third, there is no single model that is best for all situations. Even strong backbones exhibit complementary failure modes, implying that one model may be better at handling thin walls while another may be better at handling trabeculations or thrombus boundaries. This means that it is preferable to combine models rather than rely on a single architecture [4].

In this paper, we introduce a unified framework that combines *ensemble deep learning*, *federated learning*, and *genetic-algorithm (GA) fusion* for robust, privacy-preserving cardiovascular segmentation:

1. Ensemble segmentation across different backbones that work well together: We train several networks (for example, 2D/3D U-Net variants with different receptive fields, attention modules, and temporal/multi-view fusion) to take advantage of their strengths and make them more stable in difficult areas (such as the apex/base and valve planes) and across modalities.
2. Federated learning (FL) for training that keeps your privacy safe. Instead of placing all the data in a single location, participating sites train independently and only share model updates. This directly addresses privacy concerns raised in recent surveys and improves cross-site generalisation by exposing the global model to greater diversity without sharing raw images.
   Accordingly, this work focuses on methodological validation under simulated federated conditions, with large-scale real-world multi-institutional evaluation left for future investigation.
3. Instead of naive averaging, use GA-based fusion. We frame model fusion as an optimisation problem and employ a genetic algorithm (GA) to evolve fusion weights and/or model selection masks that

enhance validation performance according to task-specific metrics (Dice/HD and, when applicable, clinical indices such as EF or diameters). The evolutionary search retains specialist behaviour when it helps (e.g., in cases of a thin RV wall, small-calibre coronary arteries, and thrombus interfaces) and reduces it elsewhere.

The remainder of this paper is organized as follows. Section 2 reviews related work in cardiovascular MRI segmentation, ensemble learning, and federated optimization frameworks. Section 3 presents the proposed methodology, including the deep learning segmentation architectures, probability-level ensemble strategy, genetic algorithm-based weight optimization, and federated learning formulations. Section 4 describes the experimental setup, dataset characteristics, evaluation metrics, and implementation details. Section 5 reports and discusses the experimental results, providing a comparative analysis of centralized, ensemble, optimized, and federated scenarios. Finally, Section 8 concludes the paper and outlines potential directions for future research.

2 Related Work

This section reviews prior work on cardiovascular image segmentation, with a focus on cardiac cine-MRI. We begin by summarising traditional and machine-learning methods that use explicit anatomical priors and handcrafted representations. Next, we discuss deep learning methods, ranging from U-Net families to automated configuration systems and Transformer-based designs. We also discuss problems such as basal/apical ambiguity and cross-centre generalisation. Finally, we examine research on reliability, including quantifying uncertainty, identifying failures, and correcting mistakes with human assistance, which are becoming increasingly important for clinical use.

2.1 Classical and Machine-Learning Cardiac Segmentation Methods

When annotated datasets were scarce, classical cardiac segmentation research relied heavily on clear modelling of the heart's anatomy and the physics of image formation. In the past, early pipelines typically included preprocessing steps such as normalisation, bias correction, and ROI localisation, as well as engineered cues such as edges, region statistics, or vesselness, and constrained optimisation to ensure that the anatomy was plausible. Surveys and comparative studies have shown that cine-MRI has common limitations, including weak basal borders, partial-volume effects at the apex, contamination by papillary muscles and trabeculations, and a thin RV wall [3]. In the same context, although the field later moved toward deep learning, ideas about measuring overlap and boundaries, as well as carefully examining basal and apical slices, remain very important in today's benchmarks.

Kass et al. introduced active contours (snakes), which provide a framework for minimising energy that balances internal smoothness with image forces [8]. This resulted in smooth endocardial and epicardial contours, even amidst noisy gradients; however, edge-based forces frequently proved inadequate in areas characterised by intensity inhomogeneity or feeble boundaries. Gradient Vector Flow (GVF) also increased the capture range and made convergence better in concave areas, which is especially important for trabeculated myocardium and complex RV geometry [9]. On the other hand, region-based active contours used statistics from both the inside and outside to improve edge reliability when they were unreliable [10]. Level-set formulations enhanced numerical stability, accommodated topological modifications seamlessly, and facilitated the splitting and merging of surfaces as required [11]. Although these models were flexible, they were still affected by parameter settings and initialisation. This led to more principled priors.

To more clearly incorporate anatomical knowledge, statistical priors were added to constrain solutions to reasonable cardiac shapes and to compensate for areas that were unlearnable for fitting, using both geometric. Active Shape Models (ASMs) acquired principal modes of variation from labelled datasets and

implemented regularised fits in the presence of weak edges [12]. Active Appearance Models (AAMs) are built on ASMs by modelling both shape and texture simultaneously. This enabled fitting to use both geometric and intensity patterns [13]. Temporal and 3D AAM variants facilitated cine consistency and enhanced tracking throughout the cardiac cycle [14]. Conversely, these models degraded in the absence of pathology or acquisition characteristics within the training distribution.

Atlas-based segmentation addressed some of these problems by ensuring anatomical consistency across the world. Single-atlas methods flexibly registered a labelled template to the target and spread the labels through the transformation [15]. Similarly, multi-atlas fusion improved stability by combining labels from multiple registered atlases, thereby reducing bias toward any single atlas [16]. These methods clarified local confusion, such as distinguishing between myocardium and papillary muscles. However, they were computationally intensive and performed poorly when registration failed, especially near the base, where anatomy ends and valve motion makes alignment more challenging.

Parametric shape models utilising compact Fourier descriptors provided an alternative method for regularising segmentation. The Fourier-based representation by Staib and Duncan generated smooth closed curves governed by a limited number of parameters [17]. This compactness made it easier to couple multiple surfaces (endo/epi) in the same way, but near-elliptical assumptions may not apply to remodelled ventricles unless higher-order terms are included.

Other traditional deformable models integrated different types of information, such as combining gradient and region data in a single energy [18]. In the same context, these hybrid energies made it less sensitive to noisy edges, and region statistics stopped contour leakage across low-contrast boundaries. However, balancing these terms was challenging, and excessive regularisation could wash out thin anatomical structures such as the RV wall.

Xu and Prince [9] built on GVF to show that stronger external forces pull contours into deep concavities and make their capture range wider from rough initialisations. In the same situation, this meant that there were fewer interactions with users, but it still needed careful tuning and didn't completely clear up the epicardium-lung confusion. Later, Paragios [19] used region-driven terms and GVF-like forces to create coupled level-set evolutions for the endocardium and epicardium. This made it clear how their relative positions were limited. On the other hand, this coupling prevented spurious overlaps and separations, but strong regularisation could still smooth out fine structures.

Knowledge-based registration-and-segmentation pipelines further improved model guidance by aligning a statistical reference, typically a mean signed-distance map, with the target prior to level-set refinement [20]. In the same situation, this alignment initially accounted for differences in pose and scale, but strong priors could have biased the solutions toward average shapes when the patient's anatomy was very different.

Finite-element models (FEM) added biomechanical limits to purely geometric priors. Papademetris et al. [21] put tissue mechanics right into segmentation so that the way the surface moved followed the properties of the material. In the same context, these models naturally enforced temporal smoothness and enabled strain analysis, but they also introduced additional parameters (e.g., stiffness and anisotropy) that were difficult to estimate. Pham et al. [22] also suggested a FEM-based deformable model for 3D segmentation and tracking in cine MRI, combining spatial regularisation with temporal propagation. On the other hand, these numerical schemes required careful discretisation to remain stable and to run for long times during full cardiac cycles.

Yezzi et al.'s region-driven formulations [10] were also important because they were among the first region-based active contour formulations that combined homogeneity assumptions with boundary attraction. In the same context, using interior/exterior statistics helped stop leaks at weak borders, but it was still difficult to identify strong descriptors when intensity was unstable.

Probabilistic atlases that combined EM-based classification with Markov Random Fields further strengthened spatial regularisation. Lorenzo-Valdés et al. [15] showed that 4D probability maps that change over time can aid temporal segmentation. In this context, the quality and alignment of the atlas remained very important. Errors at the base, where the atrial inflow cuts off the ventricle, needed special heuristics. Lötjönen et al. [16] expanded this idea by using a multi-structure statistical model that included the atria, ventricles, and epicardium. This made it easier to consistently segment across structures and slices.

Finally, techniques that combined deformable registration and geometric optimisation sought to stabilise weak boundaries. Jolly et al. [23] suggested linking registration across phases with minimum-surface formulations to strengthen weak borders prior to per-frame refinement. In the same vein, extracting the shortest path in polar coordinates accelerated finding the LV cavity when it was assumed to be near-circular, but it struggled with heavily remodelled anatomies.

More recently, a different line of research has looked at traditional machine-learning classifiers as lightweight, easy-to-understand alternatives or additions to deep models. In the same vein, several studies from 2025 compared classical algorithms such as SVM, KNN, Decision Trees, AdaBoost, Random Forest, Extra Trees, CatBoost, and Gradient Boosting, highlighting their strengths in robustness, interpretability, and computational efficiency [24]. On the other hand, cardiovascular-specific pipelines that combine image-derived features (like GLCM texture descriptors) with AdaBoost-DT or KNN have done very well at predicting heart disease, often beating neural models in small samples [25]. In a similar vein, comparisons of supervised learning methods such as kNN, SVM, neural networks, and AdaBoosted decision trees on biomedical datasets—including monitoring foetal cardiac health—show that classical machine learning remains competitive when interpretability and limited data are important factors [26]. In addition, recent IEEE reports emphasise that traditional ML algorithms such as SVMs, k-NN, logistic regression, and decision trees remain useful in settings with diverse data and limited resources [27]. These findings demonstrate that conventional machine-learning techniques remain methodologically relevant in cardiovascular analysis pipelines, serving either as independent classifiers or as integral elements within hybrid segmentation-diagnosis frameworks.

Deep learning is now the dominant approach to cardiac segmentation, but many ideas from earlier methods remain valuable. These include anatomy-aware coupling, explicit shape priors, population-based atlases, region-driven energies, and temporal regularisation. These concepts continue to impact contemporary architectures, loss functions, and evaluation protocols.

2.2 Deep Learning for Cardiac Segmentation

Deep learning has transformed cardiac segmentation by using hierarchical representations learned directly from data rather than handcrafted features. Encoder–decoder architectures, such as U-Net, introduced the now-standard skip-connected design that preserves fine endocardial boundaries while retaining global contextual features [28]. As volumetric imaging became more common, extensions such as 3D U-Net improved through-plane coherence in CT and CMR volumes. However, the cost of memory and the variability of slice gaps often made 2D or 2.5D approaches better for cine MRI [29]. V-Net improved volumetric segmentation further by adding a Dice-based loss to mitigate extreme foreground imbalance and emphasising the importance of strong augmentation for stability [30]. nnU-Net was built on these ideas by

automating the whole process, from preprocessing to model capacity, patch size, and post-processing. It also demonstrated strong performance across a wide range of cardiac datasets [31].

Soon after, large-scale applications followed. Bai et al. showed that fully convolutional networks can accurately separate the LV, RV, and myocardium in cine CMR and get reliable volumetric indices, even though it is still hard to do so in basal and apical slices [32]. The ACDC challenge by Bernard et al. made public benchmarking official, enabling systematic comparison of U-Net and V-Net variants, loss functions, and pre- and post-processing strategies [6]. Later works used joint motion-segmentation learning for cine MRI to address temporal issues, employing recurrent units or optical-flow guidance to improve ED/ES alignment [33]. Complementary initiatives proposed spatial propagation strategies that deliver reliable mid-ventricular data to apical and basal slices, thereby enhancing 3D consistency across stacks acquired with anisotropic spacing [34]. Other studies improved encoder representations by using multi-scale DenseNet-based backbones and dual-loss optimisation. This made them more robust to vendor variability and helped with downstream disease classification [35]. Anatomically constrained neural networks used shape priors and attention mechanisms to make masks that didn't make sense less likely and build trust in the clinical setting [36].

Afterwards, transformer-based architectures sought to capture long-range dependencies and global structure. TransUNet and Swin-UNet are two models that combine convolutional stems with attention mechanisms. This made them perform better in hard-to-reach areas, such as the RV free wall and the basal plane [37,38]. Hybrid methods combined Transformer reasoning with nnU-Net-style pipeline automation to find a good balance between accuracy and reproducibility [39]. At the same time, several studies have proposed multi-scale supervision, feature aggregation, and boundary-aware losses to enhance performance on thin or low-contrast structures.

Deep learning in echocardiography followed a similar path, but it had to address problems specific to the modality, such as speckle noise, shadowing, and varying acoustic windows. Two-stage pipelines that first use detectors such as YOLOv7 to detect chambers and then use U-Net variants to separate LVendo, LVepi, and LA have achieved high Dice scores on CAMUS data [40]. The CAMUS dataset enabled thorough testing of deep models across AP2 and AP4 views and ED/ES phases, revealing issues with view variability and label incoherence [41]. Attention-based models such as PLANet have made it easier to capture context and maintain pixel-level label consistency in 2D echo. Variants such as MFP-UNet demonstrated improved feature aggregation across decoder layers, leading to more accurate LV delineation and clinical index estimation [42,43]. Cardiac-SegNet and other multi-structure echo systems segment the LV and LA together to improve EF estimation and anatomical consistency [44].

Deep learning has been widely used for vascular segmentation in CT and CTA, not just for chambers. Multi-scale supervised networks enhanced coronary lumen extraction [45], whereas graph convolutional networks applied to coronary surface meshes reinforced tubular topology and yielded anatomically plausible centerlines [46]. Deep models for separating calcified areas in the intracranial carotid artery were as accurate as expert readers on non-contrast CT scans and were strongly linked to stroke risk in population studies [47]. U-Net-based methods for segmenting the thoracic aorta in CT also showed high agreement with reader measurements and reliably detected aneurysms. This underscores the importance of using clinically relevant metrics, such as diameter accuracy, to evaluate segmentation [48]. Follow-up systems further improved diameter estimation by leveraging multi-view learning and Transformer-based context modelling.

In 4D Flow MRI, segmentation is the first step in analysing blood flow. Fully automated aortic masks have facilitated the accurate calculation of wall shear stress and flow-related biomarkers [49]. Extensive cohort studies utilising deep segmentation of the thoracic aorta have demonstrated that imaging phenotypes derived from automated masks can be incorporated into genetic analyses, highlighting the necessity of

uniform segmentation in population-level research [50]. Later contributions focused on cycle-aware regularisation, temporal consistency, and propagation strategies to maintain segmentation stability throughout the cardiac cycle in 4D Flow MRI [51].

Recent advancements have extended cardiac segmentation into multi-task and multi-modal domains. SURFR-Net integrates super-resolution and segmentation to enhance boundary clarity and flow quantification in 4D Flow MRI within a single framework [52]. Other systems add edge-aware and context-aware pathways to make LV/RV/myocardium segmentation on cine CMR better [53]. Directly segmenting the left ventricle from raw 4D Flow MRI without using registered cine images enables simultaneous volume and blood flow measurements and simplifies clinical workflows [54]. Comprehensive reviews of 4D Flow imaging underscore that deep learning–based segmentation is essential for the reproducible estimation of advanced haemodynamic indices [55]. All of these changes point to a clear shift away from separate encoder-decoder models and toward hybrid, anatomically aware, multi-task systems that can combine spatial, temporal, and physical constraints across different types of cardiovascular imaging.

Previous studies have independently explored ensemble and federated learning for medical image segmentation. Ensemble methods are commonly applied to improve robustness by combining multiple neural networks, while federated learning focuses on privacy preserving optimization across decentralized data sources. However, most federated segmentation studies rely on single architecture models and do not investigate architectural complementarity.

In contrast, this work integrates probability level ensemble learning with genetic algorithm based weight optimization and embeds the resulting model in a federated learning framework. While these components exist individually, their systematic combination and evaluation for cardiac MRI segmentation remain limited.

3 Proposed Approach for Cardiovascular MRI Segmentation

Cardiovascular magnetic resonance imaging (MRI) segmentation is a key part of modern cardiac image analysis. It enables quantitative measurement of ventricular volumes, myocardial mass, and functional cardiac parameters. Deep learning-based automated segmentation systems have achieved strong performance in controlled benchmark settings. However, putting this into practice in the real world remains challenging due to heterogeneous data, a lack of annotated datasets, domain shifts across institutions, and strict privacy laws that prohibit central data sharing. Given these limits, we need to develop robust, privacy-protecting, and generalisable segmentation frameworks that operate in distributed environments.

To tackle these issues, we propose a comprehensive hybrid framework that combines deep segmentation architectures, ensemble learning at the probability level, weight optimisation using genetic algorithms, and federated learning strategies. The goal is to improve segmentation reliability while preserving data privacy across institutions. The proposed pipeline systematically integrates architectural diversity with distributed optimisation to mitigate overfitting, reduce model variance, and improve generalisation across diverse clinical data distributions.

The proposed framework employs a structured pipeline with interconnected stages that improve segmentation performance over time. To ensure that the input is always the same, the MRI data are first normalised and spatially standardised. Then, several deep segmentation architectures are trained separately to get different feature representations. Using a genetic algorithm, the best weight configuration for probability-level hybrid ensemble learning is found. This combines the outputs of these models. Lastly, the optimised hybrid architecture is employed in a federated learning setting using various distributed optimisation methods.

This hierarchical structure ensures that model diversity is used locally first, and then it is incorporated into a global training strategy that protects privacy. Fig. 1 and Algorithm 1 illustrate the complete workflow.

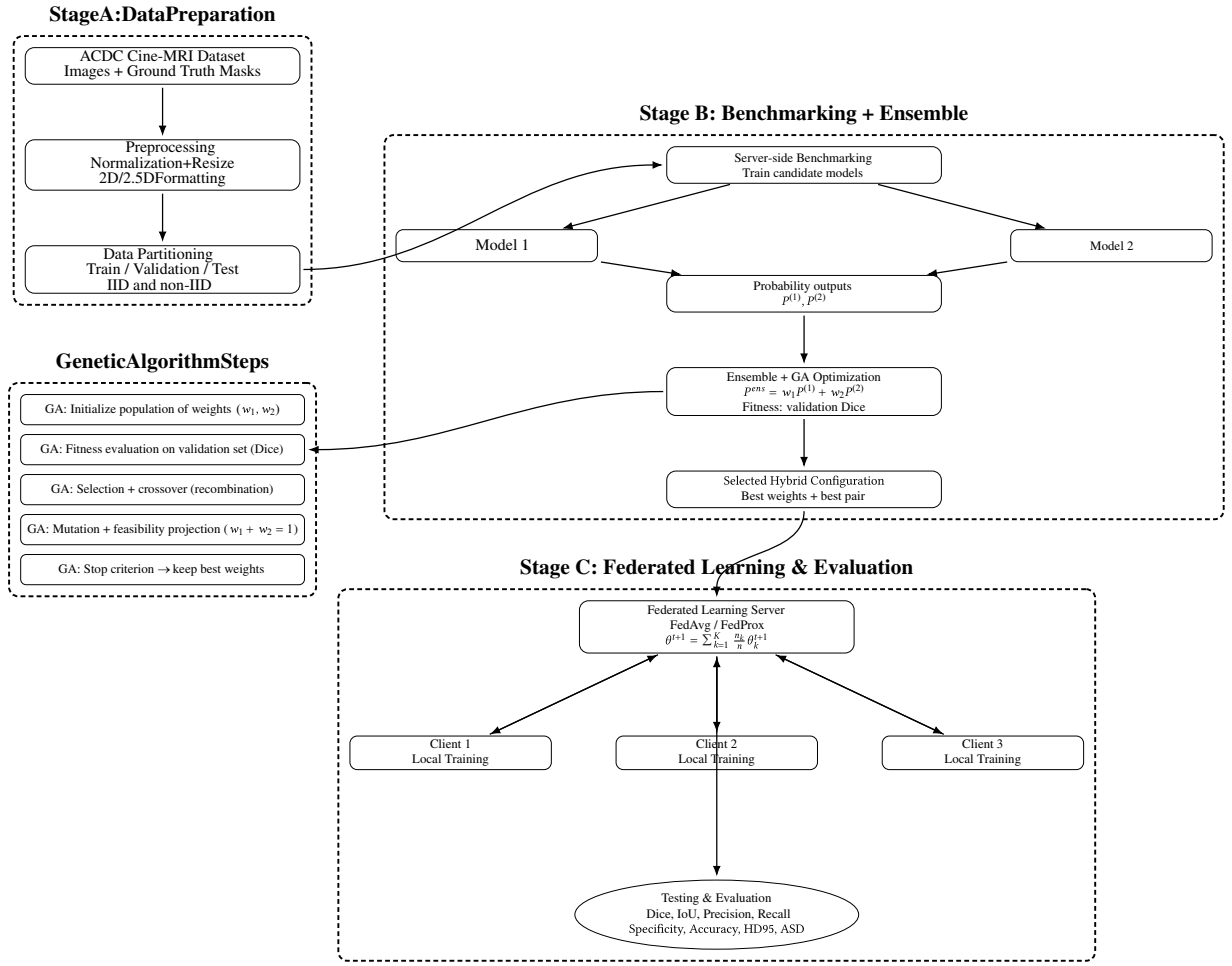


Figure 1: Diagram of the proposed cardiovascular MRI segmentation framework with the requested layout. Stage A (left) prepares and partitions the ACDC dataset. Stage B (right) benchmarks two candidate models, produces probability maps, and performs ensemble learning with GA-based optimization. The internal GA operations (initialization, fitness evaluation, crossover, mutation, and termination) are shown in a dedicated block below Stage A. Stage C (to the right of the GA block) performs federated learning (FedAvg/FedProx) using the selected hybrid configuration and evaluates the global model on held-out test data.

Algorithm 1: GA-optimized hybrid ensemble with federated learning for cardiac MRI segmentation

Require: Two pre-trained segmentation models $f_1(\cdot;\theta_1)$ and $f_2(\cdot;\theta_2)$; validation set $\mathcal{D}_{val}$; clients $\{1,\ldots,K\}$ with local datasets $\mathcal{D}_k$; federated rounds T, local epochs E.

Ensure: Fixed ensemble weights $(w_1^\star, w_2^\star)$ and trained global ensemble model $\Theta^{(T)}$.

Stage 1: Offline Ensemble Weight Optimization (Server-Side)

1: Freeze parameters θ_1 and θ_2

2: Initialize population of weight vectors $\mathbf{w} = (w_1, w_2)$ with $w_1 + w_2 = 1$

3: **for** generation $g = 1$ to G **do**

4: **for** each candidate $\mathbf{w}^{(p)}$ **do**

(Continued)

Algorithm 1 (continued)

Compute ensemble prediction:
$P_{\text{ens}}(x) = w_1^{(p)} f_1(x;\theta_1) + w_2^{(p)} f_2(x;\theta_2)$
Decode $\hat{y} = \arg\max_c P_{\text{ens}}^c(x)$
Evaluate fitness on $\mathcal{D}_{val}$
end for
Apply selection, crossover, and mutation
end for
Select optimal weights:
$(w_1^\star, w_2^\star) = \arg\max_{\mathbf{w}} \text{Fitness}(\mathbf{w})$
Stage 2: Federated Learning with Fixed Ensemble
Initialize ensemble model parameters $\Theta^{(0)}$ at server
for round $t = 0$ to $T-1$ **do**
Server broadcasts $\Theta^{(t)}$ to all clients
for all clients k **in parallel do**
$\Theta_k^{(t)} \leftarrow \Theta^{(t)}$
for local epoch $e = 1$ to E **do**
Compute ensemble-based prediction:
$P_{\text{ens}}(x;\Theta_k) = w_1^\star f_1(x;\Theta_k) + w_2^\star f_2(x;\Theta_k)$
Update Θ_k using segmentation loss
end for
Send updated $\Theta_k^{(t+1)}$ to server
end for
Aggregate:
$\Theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} \Theta_k^{(t+1)}$
end for
return Final global ensemble model $\Theta^{(T)}$

3.1 Dataset Description and Preprocessing

The Automatic Cardiac Diagnosis Challenge (ACDC) dataset [6] consists of short-axis cine-MRI volumes acquired from multiple patients with varying pathological conditions. Each case includes expert-annotated masks for the right ventricle (RV), myocardium (MYO), and left ventricle (LV).

As shown in Fig. 2, this dataset comprises 2D cine MR images from 100 patients acquired on various 1.5 and 3T MR scanners at different temporal resolutions. Manual segmentations are provided for the end-diastolic (ED) and end-systolic (ES) cardiac phases for right ventricle (RV), left ventricle (LV) and myocardium (MYO).

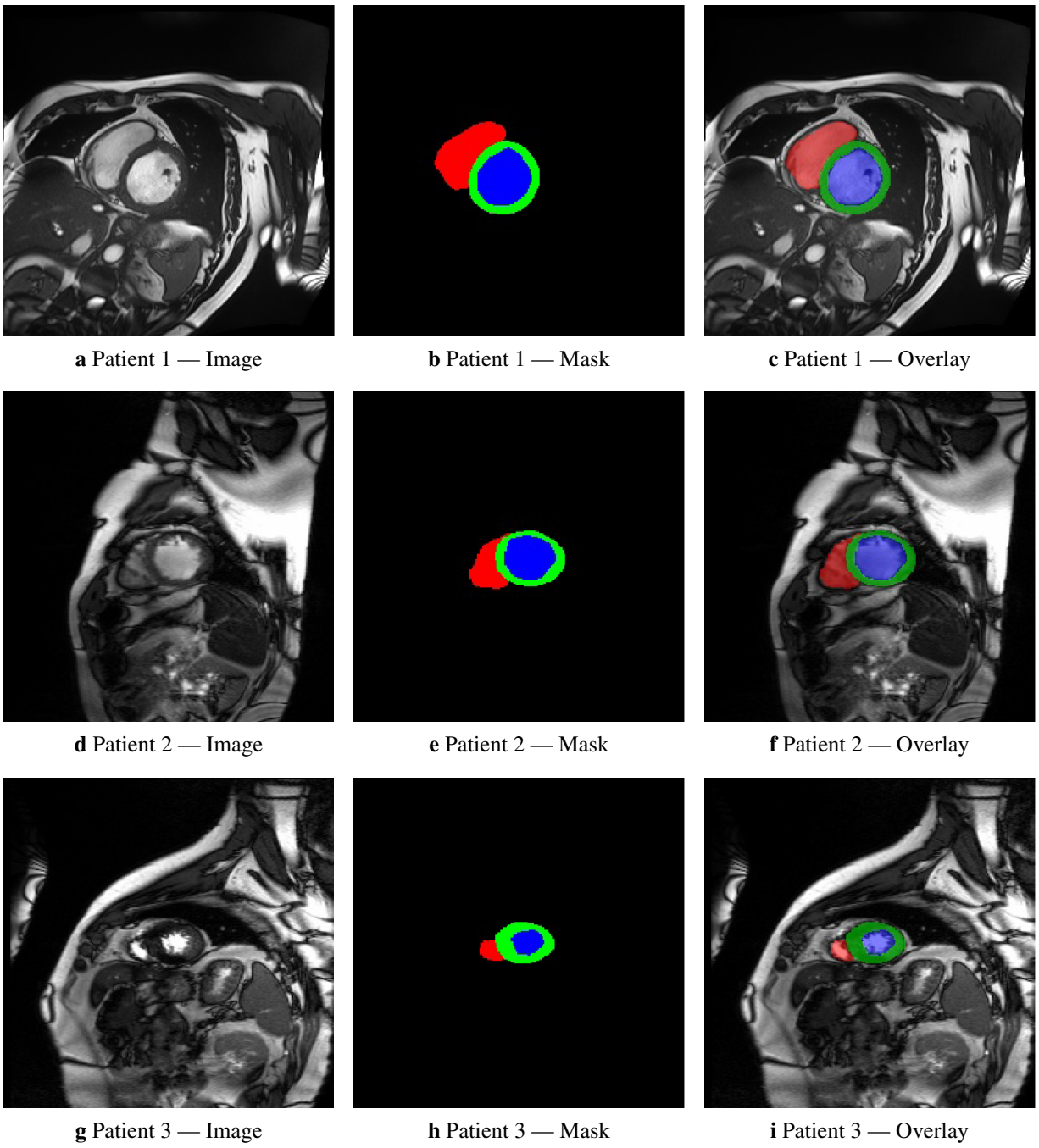
The dataset represents realistic clinical variability, including differences in anatomy, pathology, and acquisition settings. Each MRI volume is defined as:

$$\mathcal{V} = \{I_s\}_{s=1}^{S} \tag{1}$$

where $I_s \in \mathbb{R}^{H\times W}$ is a 2D slice.

The corresponding mask is:

$$Y_s \in \{0,1,2,3\}^{H\times W} \tag{2}$$



a Patient 1 — Image **b** Patient 1 — Mask **c** Patient 1 — Overlay

d Patient 2 — Image **e** Patient 2 — Mask **f** Patient 2 — Overlay

g Patient 3 — Image **h** Patient 3 — Mask **i** Patient 3 — Overlay

Figure 2: Per-patient rows showing (Col-1) original image, (Col-2) segmentation mask, and (Col-3) overlay from train dataset. Red, green, and blue correspond to the right ventricle (RV), left ventricular myocardium (MYO), and left ventricular cavity (LV), respectively, following the standard ACDC label encoding.

To ensure numerical stability and improve convergence, each slice is normalized using z-score normalization:

$$\tilde{I}_s = \frac{I_s - \mu_s}{\sigma_s + \epsilon} \tag{3}$$

This normalization reduces inter-scan intensity variability, which is common in MRI acquisitions. Images are then resized to a fixed spatial resolution:

$$I'_s \in \mathbb{R}^{256\times256} \tag{4}$$

For 2.5D models, contextual slices are concatenated:

$$I_s^{25D} = [I_{s-1}, I_s, I_{s+1}] \tag{5}$$

All ACDC cases used in this study include complete expert annotations, and no missing segmentation labels were observed. The dataset was used without case- or slice-level exclusion based on anatomical appearance. As a result, challenging regions such as basal and apical slices, which exhibit higher anatomical variability and ambiguity, were retained. This ensures that model evaluation reflects realistic clinical conditions rather than a curated or simplified subset.

3.2 Deep Learning Models for Segmentation

Cardiac segmentation is formulated as a multi-class semantic segmentation task. The objective is to learn a nonlinear mapping:

$$f_\theta : I \to P \tag{6}$$

where:

- θ denotes the learnable model parameters,
- $P \in \mathbb{R}^{C\times H\times W}$ represents the predicted probability maps,
- $C = 4$ corresponds to background and three anatomical classes (RV, MYO, LV).

The predicted segmentation map is computed as:

$$\hat{Y}(x) = \arg\max_c P_c(x) \tag{7}$$

where $\hat{Y}(x)$ denotes the predicted class label at pixel location x.

To increase architectural diversity and reduce model bias, we evaluate eight deep segmentation models, each capturing complementary spatial and contextual features to enhance robustness and generalization.

A combined cross-entropy and Dice loss was employed to balance pixel-wise classification accuracy and region-level overlap optimization. In our experiments, cross-entropy contributes to stable, smooth convergence during early training by providing dense pixel-level supervision, particularly important in the federated setting where local updates are performed on limited client data. Dice loss complements this by directly optimizing region overlap and mitigating class imbalance, especially for anatomically thin and variable structures such as the myocardium and right ventricle.

The combination of these loss terms was observed to improve convergence stability across clients and reduce sensitivity to local data imbalance, thereby facilitating more consistent aggregation during federated optimization. Moreover, by emphasizing both pixel-level and region-level objectives, the loss formulation supports robust ensemble fusion, as individual model predictions exhibit improved structural consistency, particularly at object boundaries.

3.2.1 UNet++ (EfficientNet-B3 and EfficientNet-B4)

UNet++ [56] is a more advanced version of the classic U-Net. It adds nested and densely connected skip pathways to the original design. The main goal of UNet++ is to close the semantic gap between encoder and decoder feature maps. This is particularly important in medical image segmentation, where precise boundaries are required. To get the right shape of the myocardial borders and ventricular cavities in cardiovascular MRI segmentation, you need to keep improving the spatial details at different resolution levels.

Let $X^{i,j}$ denote the feature map at depth i and stage j within the nested decoder structure. The dense skip connection mechanism is defined as:

$$X^{i,j} = H\left(\left[X^{i,j-1}, U(X^{i+1,j-1})\right]\right) \tag{8}$$

where:

- $H(\cdot)$ represents a convolutional transformation,
- $U(\cdot)$ denotes upsampling,
- $[\cdot]$ indicates concatenation.

This nested architecture facilitates incremental feature improvements and makes gradient propagation during training smoother. UNet++ differs from the standard U-Net by incorporating additional dense connections that enable supervision at multiple levels. This makes convergence more stable and segmentation more accurate.

The encoder backbone is based on EfficientNet-B3 and EfficientNet-B4, which use compound scaling to balance network depth, width, and input resolution. The formula for compound scaling is:

$$d = \alpha^{\phi}, \quad w = \beta^{\phi}, \quad r = \gamma^{\phi} \tag{9}$$

subject to:

$$\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \tag{10}$$

where:

- d is network depth,
- w is network width,
- r is input resolution,
- ϕ is a scaling coefficient.

EfficientNet-B3 and B4 provide strong feature extraction capabilities while maintaining computational efficiency. In cardiac MRI segmentation, these encoders capture both global anatomical structures and fine myocardial textures.

The training objective combines cross-entropy and Dice loss to balance pixel-wise classification and overlap optimization:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{Dice} \tag{11}$$

where Dice loss is defined as:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i} \tag{12}$$

UNet++ with EfficientNet backbones provides strong boundary refinement, improved structural consistency, and enhanced robustness to intensity variability in cardiac MRI data.

3.2.2 FPN (EfficientNet-B3 and EfficientNet-B4)

The Feature Pyramid Network (FPN) [57] is made to use multi-scale feature representations through a top-down pathway with lateral connections. When you do cardiovascular MRI segmentation, anatomical structures like the myocardium and ventricles show up at different scales in space. For accurate segmentation, it's important to get both the big picture and the small details.

Let C_l denote the feature map extracted at level l from the encoder. The top-down feature fusion mechanism is defined as:

$$P_l = Conv\left(C_l + Up(P_{l+1})\right) \tag{13}$$

where:

- P_l is the pyramid feature map at level l,
- $Up(\cdot)$ denotes upsampling,
- $Conv(\cdot)$ represents convolution.

This recursive feature aggregation integrates high-level semantic features with low-level spatial details. The pyramid representation improves the detection of thin myocardial walls while preserving the overall structure of the ventricular cavities.

Like UNet++, FPN uses EfficientNet-B3 and EfficientNet-B4 as backbone encoders. The pyramid structure of EfficientNet combines hierarchical features extracted over time, enabling the learning of strong multi-resolution representations.

The final segmentation probability map is computed as:

$$P = Softmax(Conv(P_2)) \tag{14}$$

where P_2 corresponds to the highest-resolution pyramid feature map.

FPN is particularly effective for segmenting cardiac MRIs because it scales well. It improves segmentation in both basal and apical slices, where the anatomical shapes differ markedly. Also, the pyramid aggregation method makes the system more robust to partial-volume effects and low-contrast boundaries that are common in cine-MRI sequences.

FPN focuses on hierarchical semantic consistency across scales, making it well-suited to the UNet++ ensemble framework. This differs from architectures that use only decoders.

3.2.3 DeepLabV3+

DeepLabV3+ [58] is a strong semantic segmentation architecture that uses atrous convolutions and an encoder-decoder refinement to capture both global context and fine structural details. In cardiovascular MRI segmentation, accurately defining the boundaries of the myocardium and the ventricular cavities requires large receptive fields to model the overall shape while maintaining spatial accuracy at anatomical edges. DeepLabV3+ meets this need by using Atrous Spatial Pyramid Pooling (ASPP), which uses parallel dilated convolutions with different dilation rates to get multi-scale contextual features.

Dilated convolution is defined as:

$$y[i] = \sum_{k} x[i + r \cdot k]w[k] \tag{15}$$

where r is the dilation rate. By increasing r, the receptive field expands without increasing the number of parameters or reducing feature map resolution. The ASPP module aggregates features across multiple dilation rates:

$$ASPP(x) = \sum_{r \in R} Conv_r(x) \tag{16}$$

This mechanism enables the model to capture both fine myocardial textures and larger ventricular structures simultaneously. The decoder module also improves segmentation by combining low-level spatial features from earlier encoder layers, thereby making boundaries more precise. DeepLabV3+ performs particularly well in the presence of substantial anatomical variability because its multi-scale representation makes it less sensitive to structural and disease-related changes.

3.2.4 UNet 2.5D (EfficientNet-B3)

2D convolutional networks [59] process slices separately, and 3D networks need a lot of computing power. The 2.5D approach is a good middle ground. In cardiovascular MRI, adjacent slices are very similar because the anatomy is continuous along the short-axis direction. Pure 2D models may not exploit this inter-slice dependency, potentially resulting in non-smooth segmentation across slices.

The 2.5D input representation is constructed as:

$$I_s^{25D} = [I_{s-1}, I_s, I_{s+1}] \tag{17}$$

This provides contextual information from nearby slices while remaining as fast as 2D models. The EfficientNet-B3 encoder takes this three-channel input and extracts hierarchical spatial features. This allows the network to model small changes in myocardial thickness and ventricular geometry across slices. The 2.5D UNet architecture enhances structural coherence and diminishes slice-to-slice segmentation discrepancies by integrating constrained 3D context without comprehensive volumetric convolution, particularly in apical and basal areas characterised by rapid anatomical variation.

3.2.5 DSE-Net

DSE-Net (Deep Spatial Encoding Network) [60] improves segmentation performance by encoding spatial features and combining them hierarchically. In cardiac MRI, accurate segmentation is challenging due to unclear boundaries and low contrast between the myocardium and surrounding tissues. DSE-Net addresses this problem by using learnable spatial-encoding weights to combine features from multiple encoder levels.

Let encoder features be:

$$F = \{F_1, F_2, F_3, F_4\} \tag{18}$$

Spatial aggregation is defined as:

$$S = \sum_{i=1}^{4} W_i * F_i \tag{19}$$

where W_i are learnable convolutional kernels. This weighted aggregation mechanism allows the model to emphasize discriminative spatial cues at different scales. By combining shallow high-resolution features with deeper semantic representations, DSE-Net improves boundary continuity and anatomical consistency. Furthermore, the spatial encoding mechanism enhances robustness against intensity inhomogeneity and motion artefacts commonly observed in cardiac cine-MRI sequences. As a result, DSE-Net contributes complementary predictions within the ensemble framework.

3.3 Ensemble Learning

Deep segmentation models frequently demonstrate synergistic strengths. UNet++ excels at refining boundaries with dense skip connections, FPN excels at capturing multi-scale hierarchical representations, and DeepLabV3+ improves contextual understanding with atrous spatial pyramid pooling. Nonetheless, no single architecture consistently prevails across anatomical structures and patient variability. Differences in myocardial thickness, ventricular geometry, and imaging conditions can make some models perform less well in certain areas. This encourages the application of ensemble learning to integrate predictions from various models, thereby leveraging their complementary attributes and enhancing overall robustness.

Given probability outputs $P^{(m)}$ from M models, the ensemble prediction is computed as:

$$P^{ens} = \sum_{m=1}^{M} w_m P^{(m)} \tag{20}$$

subject to the constraint:

$$\sum_{m=1}^{M} w_m = 1 \tag{21}$$

where w_m denotes the contribution weight of the m-th model.

By averaging outputs that are well correlated, ensemble learning reduces model variance and stabilises predictions. Hard voting is not as effective as probability-level fusion because it retains confidence information and yields smoother decision boundaries. This mixed strategy makes the model more robust to its own specific errors, reduces overfitting, and improves generalisation across data distributions that differ. Ensemble learning is especially useful for balancing boundary accuracy, structural continuity, and contextual modelling in cardiovascular MRI segmentation.

Genetic Algorithm Optimization (Extended Description)

Choosing the optimal ensemble weights is challenging because the models' outputs interact nonlinearly. We use a Genetic Algorithm (GA) to search for the best combination space, adapting to the data rather than assigning weights by hand.

The optimization objective is defined as:

$$\max_{w} \frac{1}{N} \sum_{n=1}^{N} Dice_n \tag{22}$$

where N denotes the number of validation samples and $Dice_n$ is the Dice score for sample n.

The GA begins by initializing a population of candidate weight vectors satisfying:

$$\sum_{m=1}^{M} w_m = 1, \quad w_m \geq 0 \tag{23}$$

Each candidate solution (chromosome) encodes a feasible weight configuration. The fitness of each chromosome is computed using the validation Dice score obtained from the ensemble prediction. The algorithm then performs iterative evolutionary operations:

- **Selection:** Individuals with higher fitness values are more likely to be selected for reproduction.
- **Crossover:** Weight vectors are recombined to explore new regions of the solution space.
- **Mutation:** Small random perturbations are introduced to prevent premature convergence and maintain diversity.
- **Elitism:** The best-performing solutions are preserved across generations.

The GA converges to the best weight distribution over many generations, yielding the best segmentation performance. This adaptive optimisation strategy doesn't get stuck in local minima and accounts for how models depend on one another in complex ways. The GA finds ensemble configurations that balance boundary accuracy, contextual consistency, and structural coherence by exploring the weight space globally instead of relying on manual tuning. The best hybrid configuration is chosen and then used as the base ensemble model in the federated learning framework.

Let $\{f_m(\cdot;\theta_m)\}_{m=1}^{M}$ denote a set of M pre-trained segmentation models, where θ_m represents the parameters of model m. Given an input image x, each model produces a class probability map $P_m(x)$.

The ensemble prediction is defined as

$$P_{\text{ens}}(x) = \sum_{m=1}^{M} w_m P_m(x), \tag{24}$$

subject to the constraints

$$\sum_{m=1}^{M} w_m = 1, \quad w_m \geq 0, \tag{25}$$

where w_m denotes the fixed fusion weight assigned to model m, determined during offline genetic algorithm optimization. The final segmentation label is obtained by applying the arg max operator over the ensemble class probabilities.

3.4 Federated Learning

Storing all MRI data on a single server for centralised training violates patient privacy and institutional data-sharing regulations. Federated learning addresses this problem by enabling models to be trained jointly without sharing raw data. Instead, the model parameters are shared and combined, while the data stays at each participating institution.

The global optimization objective is defined as:

$$\min_{\theta} \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta) \tag{26}$$

where:

- K denotes the number of clients,
- n_k represents the number of local samples at client k,
- $n = \sum_{k=1}^{K} n_k$ is the total number of samples,
- $\mathcal{L}_k(\theta)$ is the local loss function.

Clients use their private data to update their local models and send the new parameters to the central server during each communication round. The server combines these updates to form a new global model, which it then sends back to clients for the next training round. This process repeats until convergence.

Federated learning enables institutions to collaborate to share knowledge while preserving data privacy. This paradigm is particularly useful for cardiovascular MRI segmentation because there is limited annotated data per centre and substantial variation across patient groups and imaging devices. The proposed framework improves robustness, facilitates generalisation across diverse data distributions, and ensures compliance with medical privacy standards by combining federated optimisation with hybrid ensemble learning.

In the federated setting, each client $k \in \{1, \ldots, K\}$ optimizes the same ensemble-based segmentation model with parameters θ using its private local dataset $\mathcal{D}_k$. The local optimization objective at client k is defined as

$$\min_{\theta} \mathcal{L}_k(\theta), \tag{27}$$

where $\mathcal{L}_k(\theta)$ denotes the segmentation loss computed from the ensemble predictions on $\mathcal{D}_k$.

After local optimization, model parameters are transmitted to the server and aggregated using a standard federated optimization strategy, such as Federated Averaging (FedAvg) or proximal regularization (FedProx). Throughout federated training, the ensemble fusion weights remain fixed and are not modified during local updates or global aggregation.

3.4.1 FedAvg

Federated Averaging (FedAvg) is the main algorithm in federated learning. It enables models to be trained across multiple computers without storing raw data in a single location. In cardiovascular MRI segmentation, each client trains its own model on private MRI data and sends only the model parameters to the central server. This system ensures that medical data protection rules are followed while also enabling people to learn together. FedAvg is a useful approach for protecting privacy in multi-centre medical imaging applications because it separates data storage from model optimisation.

At communication round t, each client performs local stochastic gradient descent updates:

$$\theta_k^{t+1} = \theta^t - \eta \nabla \mathcal{L}_k(\theta^t) \tag{28}$$

where η is the learning rate and $\mathcal{L}_k$ denotes the local objective function of client k.

The server aggregates client updates as:

$$\theta^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{t+1} \tag{29}$$

where n_k is the number of samples at client k, and $n = \sum_{k=1}^{K} n_k$.

This weighted averaging ensures that clients with larger datasets have a greater impact on the global model. FedAvg works best when the data is IID and the local gradients are close to the global objective. However, in non-IID settings, which are common in medical imaging due to demographic, scanner, and acquisition-protocol differences, client updates may perform differently. Even with this limitation, FedAvg remains easy to use and scales well on modern computers, making it a strong and widely used starting point for distributed cardiovascular segmentation.

3.4.2 FedProx

FedAvg performs well with data that are roughly IID, but it may be less effective in settings that are highly heterogeneous (non-IID), which is common in multi-centre cardiovascular MRI datasets. Variability among clients can be significant due to differences in patient demographics, scanner vendors, acquisition protocols, and annotation styles. This difference may cause local updates to drift away from the global goal, thereby destabilising convergence and degrading segmentation performance.

FedProx fixes this problem by adding a proximal regularisation term to the local objective function. The main idea is to limit local model updates to keep them close to the current global model parameters. This will prevent excessive divergence caused by different gradients.

The modified local optimization objective for client k is defined as:

$$\min_{\theta_k} \mathcal{L}_k(\theta_k) + \frac{\mu}{2}\|\theta_k - \theta^t\|^2 \tag{30}$$

where:

- $\mathcal{L}_k(\theta_k)$ is the local loss function,
- θ^t represents the global model at communication round t,
- $\mu > 0$ is the proximal regularization coefficient.

The additional proximal term penalizes large deviations between the locally updated parameters θ_k and the global parameters θ^t. This regularization effectively limits the influence of extreme local gradients and encourages more consistent updates across clients.

After local optimization, the server performs weighted aggregation similarly to FedAvg:

$$\theta^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \theta_k^{t+1} \tag{31}$$

The proximal coefficient μ controls the trade-off between local adaptation and global consistency. When $\mu = 0$, FedProx reduces to standard FedAvg. As μ increases, local models are more strongly constrained to remain close to the global model, which improves stability in highly non-IID scenarios.

In cardiovascular MRI segmentation, FedProx is particularly useful when institutions involved have very different data types. FedProx improves convergence stability, reduces oscillatory behaviour, and improves generalisation performance across a wide range of imaging domains by reducing client drift. Additionally, the proximal regularisation framework guarantees theoretical convergence under bounded heterogeneity assumptions, making FedProx a strong alternative to FedAvg for distributed medical imaging applications.

3.5 Final Hybrid Federated Segmentation

The final segmentation integrates genetically optimized ensemble learning within federated training:

$$\hat{Y}(x) = \arg\max_{c} \left(\sum_{m=1}^{M} w_m P_c^{(m)}(x) \right) \tag{32}$$

This unified framework combines architectural diversity, evolutionary optimisation, and privacy-preserving distributed learning to enhance cardiovascular MRI segmentation performance in multi-centre environments.

To remove ambiguity regarding the interaction between ensemble learning and federated optimization, we explicitly clarify the temporal and functional role of each component in the proposed pipeline. Ensemble learning is first performed in a centralized and offline manner to identify complementary base models and determine fixed fusion weights using a validation set. This ensemble configuration remains constant throughout subsequent training stages.

Once the optimal ensemble configuration is determined, the resulting hybrid model serves as the base architecture in the federated learning framework. During federated training, each client trains the same fixed ensemble model on local data, and only model parameters are exchanged with the central server. Ensemble fusion is therefore not performed dynamically across clients nor adapted during communication rounds, but is embedded within the model's forward pass using fixed weights.

At inference time, the trained federated global model produces ensemble predictions through the same probability-level fusion mechanism. This design ensures consistency between centralized validation, federated optimization, and final deployment, while avoiding additional communication or synchronization overhead associated with client-wise ensemble adaptation.

3.6 Simulated Client Partitioning for Federated Learning

Federated learning experiments were conducted using a simulated multi-client setup based on the ACDC training dataset. Specifically, the data were partitioned into K = 5 clients, each representing an independent data holder. All federated experiments were performed over 20 communication rounds, with each client executing one local training epoch per round. Full client participation was adopted, meaning that all clients contributed updates during every communication round.

To evaluate the effect of data heterogeneity, both IID and non-IID client partitioning strategies were considered. For the IID setting, training cases were randomly shuffled and uniformly distributed across clients. For the non-IID setting, a patient-level partitioning strategy was employed, where each client received data from a disjoint subset of patient identifiers. This induces strong distributional heterogeneity across clients, as anatomical characteristics and slice distributions vary by patient. Client datasets were therefore imbalanced in size and composition, reflecting realistic variability in local data availability.

Although this setup does not represent a true multi-institutional cohort with scanner- or site-specific acquisition differences, it provides a controlled and reproducible platform to study federated optimization behavior under simulated heterogeneity.

4 Results and Discussion

This section provides a thorough assessment of the proposed cardiovascular MRI segmentation framework across diverse experimental settings. The evaluation encompasses centralised deep learning benchmarking, probability-level ensemble learning, genetic algorithm (GA)-driven weight optimisation, and federated learning applied to both IID and non-IID data distributions. We evaluate performance using metrics that quantify overlap, number of pixels, and boundary length. We also look into architectural complementarity and how adaptive weight optimisation affects the stability of distributed training and the strength of segmentation.

4.1 Quantitative Evaluation Metrics

This subsection describes the evaluation metrics used to measure segmentation quality and boundary accuracy. Since medical image segmentation requires both volumetric overlap precision and contour accuracy, multiple complementary metrics are employed.

4.1.1 Dice Similarity Coefficient (DSC)

The Dice coefficient measures the overlap between predicted segmentation P and ground truth G:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} \tag{33}$$

where $|P|$ and $|G|$ denote the number of foreground pixels. Dice ranges from 0 to 1, with 1 indicating perfect overlap.

4.1.2 Intersection over Union (IoU)

IoU evaluates the ratio between the intersection and the union:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|} \tag{34}$$

IoU is more penalizing than Dice for small boundary mismatches.

4.1.3 Precision and Recall

Precision quantifies over-segmentation, while recall measures under-segmentation:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{35}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{36}$$

where TP, FP, and FN denote true positives, false positives, and false negatives.

4.1.4 Pixel-Wise Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{37}$$

Although accuracy reflects overall correctness, it may be biased by large background regions.

4.1.5 95th Percentile Hausdorff Distance (HD95)

Boundary agreement is evaluated using the 95th percentile Hausdorff distance:

$$HD_{95}(P, G) = \max\left\{ \underset{p \in P}{\text{percentile}_{95}} \min_{g \in G} d(p, g), \underset{g \in G}{\text{percentile}_{95}} \min_{p \in P} d(g, p) \right\} \tag{38}$$

where $d(\cdot, \cdot)$ denotes Euclidean distance.

4.1.6 Average Surface Distance (ASD)

$$ASD(P, G) = \frac{1}{|P| + |G|} \left(\sum_{p \in P} \min_{g \in G} d(p, g) + \sum_{g \in G} \min_{p \in P} d(g, p) \right) \tag{39}$$

ASD captures the mean contour deviation and complements HD95.

4.2 Comparative Analysis of Deep Segmentation Models

This part examines how well centralised performance performs across eight different segmentation architectures. Table 1 shows a full comparison of eight segmentation architectures that were tested on the ACDC dataset. The results show clear patterns in performance across metrics measuring overlap (Dice and IoU), pixel-level metrics (precision, recall, and accuracy), and metrics sensitive to boundaries (HD95 and ASD). Overall, encoder-decoder architectures with dense multi-scale feature aggregation perform better at segmenting, particularly in capturing fine myocardial contours.

Table 1: Performance comparison of segmentation models on the ACDC dataset. Validation Dice excludes background. Best results are shown in **bold**, second-best underlined. Lower values are better for HD95 and ASD.

Model	Dice	IoU	Precision	Recall	Accuracy	HD95 (px)	ASD (px)
UNet++ (EffNet-B3)	**0.8890**	**0.8078**	0.9070	<u>0.8829</u>	**0.9976**	2.4033	**0.8213**
FPN (EffNet-B3)	<u>0.8887</u>	<u>0.8062</u>	0.8983	**0.8875**	<u>0.9976</u>	**2.3926**	<u>0.8602</u>
FPN (EffNet-B4)	0.8843	0.8006	<u>0.9133</u>	0.8679	0.9975	<u>2.4680</u>	0.8885
DeepLabV3+ (EffNet-B4)	0.8763	0.7892	0.9121	0.8568	0.9974	2.5499	0.8988
UNet++ (EffNet-B4)	0.8677	0.7820	0.9134	0.8495	0.9972	2.8304	0.9920
DeepLabV3+ (ResNet-101)	0.8485	0.7557	**0.9149**	0.8181	0.9970	3.2021	1.1576
DSE-Net	0.7134	0.6083	0.8626	0.6852	0.9950	5.3445	2.1467
UNet (EffNet-B3, 2.5D)	0.1699	0.0999	0.3494	0.2027	0.9776	83.8356	46.0210

DSE Net was included solely as a representative spatial-encoding baseline to assess the transferability of non-cardiac architectures to cine-MRI segmentation. As the model was originally designed for mobile tongue image segmentation, it is not considered a competitive cardiac-specific baseline. Its substantially inferior performance highlights the importance of domain-specific architectural design. Consequently, DSE Net predictions were excluded from the optimized ensemble and federated learning experiments and retained only for ablation analysis.

UNet++ achieved the best overlap-based performance (Dice and IoU) while maintaining competitive boundary metrics. In particular, UNet++ with EfficientNet-B3 achieved the highest Dice score (0.8890) and performed slightly better than its deeper EfficientNet-B4 counterpart on both Dice and IoU, suggesting that it generalises better. This behaviour shows that adding more encoder layers does not always improve test performance, especially when the dataset is not very large. EfficientNet-B3 appears to have sufficient representational power without introducing additional complexity that could increase the likelihood of overfitting.

Also, the nested dense skip connections in UNet++ reduce the semantic gap between encoder and decoder features. This architectural design facilitates the reconstruction of fine anatomical boundaries, particularly at ventricular interfaces. The low ASD value and good HD95 performance indicate that it preserves the contour's smoothness and the boundary's continuity. In cardiac MRI segmentation, precise boundary delineation is crucial for dependable ventricular volume estimation and myocardial thickness calculation, rendering these enhancements clinically significant.

FPN architectures, on the other hand, showed competitive but slightly lower performance. FPN effectively integrates information across scales into a single representation, but it lacks UNet++'s nested skip refinement mechanism. Although recall remains high, the boundary metrics are slightly worse than those of UNet++. This means that FPN is effective at capturing global structural information, but its feature fusion strategy may not be as effective at recovering fine myocardial edges.

Similarly, DeepLabV3+ variants achieved very high precision, particularly with the EfficientNet-B4 backbone. The Atrous Spatial Pyramid Pooling (ASPP) module increases the receptive field without reducing resolution, enabling the model to capture more contextual information. The slightly lower recall relative to UNet++ suggests that extensive contextual aggregation may smooth thin anatomical structures, leading to mild under-segmentation in challenging areas such as apical slices. Also, DeepLabV3+ with ResNet-101 achieved higher HD95 and ASD values, suggesting that encoder architecture affects boundary stability and that deeper residual networks may not always yield the best spatial accuracy for cine-MRI data.

DSE-Net exhibited markedly inferior overlap performance and significantly elevated boundary errors in comparison to encoder–decoder architectures, as illustrated in Fig. 3. Although it employs spatial encoding and attention mechanisms, the model appears less well-suited to the intensity and shape changes observed in cardiac cine sequences. The high HD95 and ASD values show that the boundary reconstruction is unstable and the contour smoothness is lower. These results indicate that spatial encoding strategies alone may be inadequate without strong multi-scale feature integration and well-optimized encoder backbones that can capture both global and local representations.

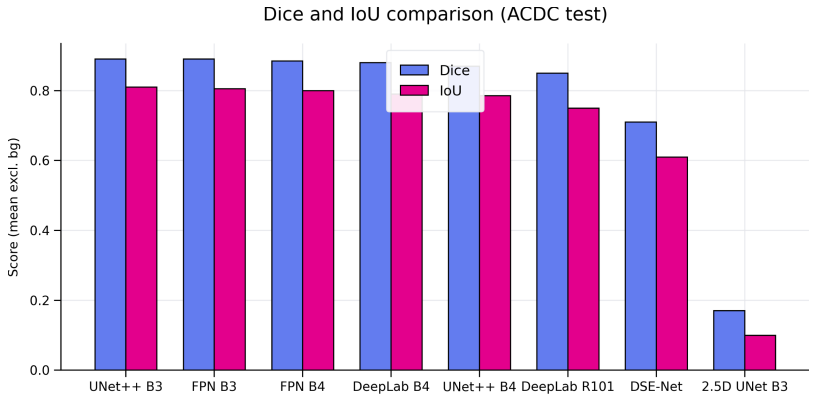


Figure 3: Compare different models by DICE and IOU metrics.

The 2.5D UNet model performed even worse. Adding adjacent slices should, in principle, provide contextual information between slices, but the results show a substantial drop in test performance and very large boundary distances. This behaviour could be due to differences in the anatomy of each slice, to the varying slice thicknesses in the ACDC dataset, and to the lack of explicit volumetric convolutional modelling. With the current training setup, concatenating neighbouring slices may propagate inconsistencies between slices rather than providing useful structural context, thereby making segmentation unstable. As a result, the 2.5D approach didn't perform well on this dataset and appears to be affected by inter-slice differences.

Metrics based on boundaries make it easier to see architectural differences. UNet++ with EfficientNet-B3 achieved the lowest ASD value and competitive HD95 performance, indicating highly accurate contours. In practice, even small changes in the boundaries can significantly affect volumetric and functional measurements. Thus, automated cardiac assessment systems require models with lower boundary errors.

In conclusion, these results show that the most effective approach for stable, accurate segmentation on ACDC is to use nested encoder-decoder architectures with efficient feature extractors. Increasing the encoder depth beyond a certain point yields diminishing returns, whereas effective multi-scale feature fusion remains very important. Architectures that strike the best balance between global contextual modelling and local boundary refinement achieve the best trade-off between precision and recall, making them more robust across diverse heart structures.

Fig. 4 shows a qualitative comparison of the segmentation outputs from all the models that were tested on a representative ACDC test slice. The figure shows the original cardiac MRI image, the ground-truth annotation (colour-coded), and the predicted segmentation maps with their overlays. As we can see, UNet++ with EfficientNet-B3 and EfficientNet-B4 produces segmentations that are very close to the true contours, especially at the ventricular edges. The predicted masks exhibit smooth shapes consistent with anatomy and change little at the edges. The FPN and DeepLabV3+ models also perform well in capturing global cardiac structures, but they exhibit small irregularities at boundaries, particularly at the myocardial interface. DSE-Net, on the other hand, does not do as well at drawing precise contours, and there are some minor shape distortions. The 2.5D UNet model exhibits the most pronounced degradation, with predictions that are fragmented and noisy outside the anatomical region of interest. This confirms the quantitative results shown in Table 1. Overall, the visual inspection corroborates the quantitative results, indicating that nested encoder–decoder architectures are better at maintaining spatial consistency and boundary accuracy in cardiac cine-MRI segmentation.

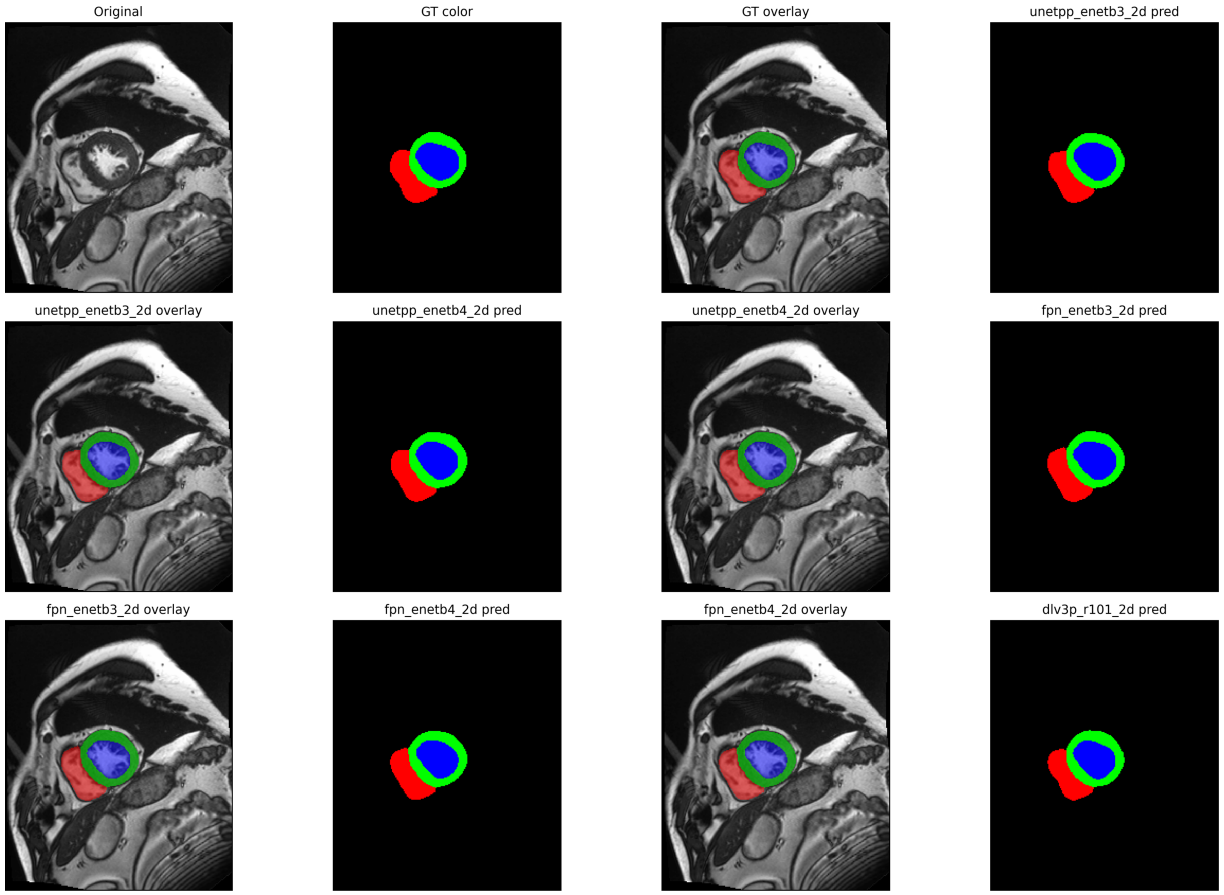


Figure 4: (Continued)

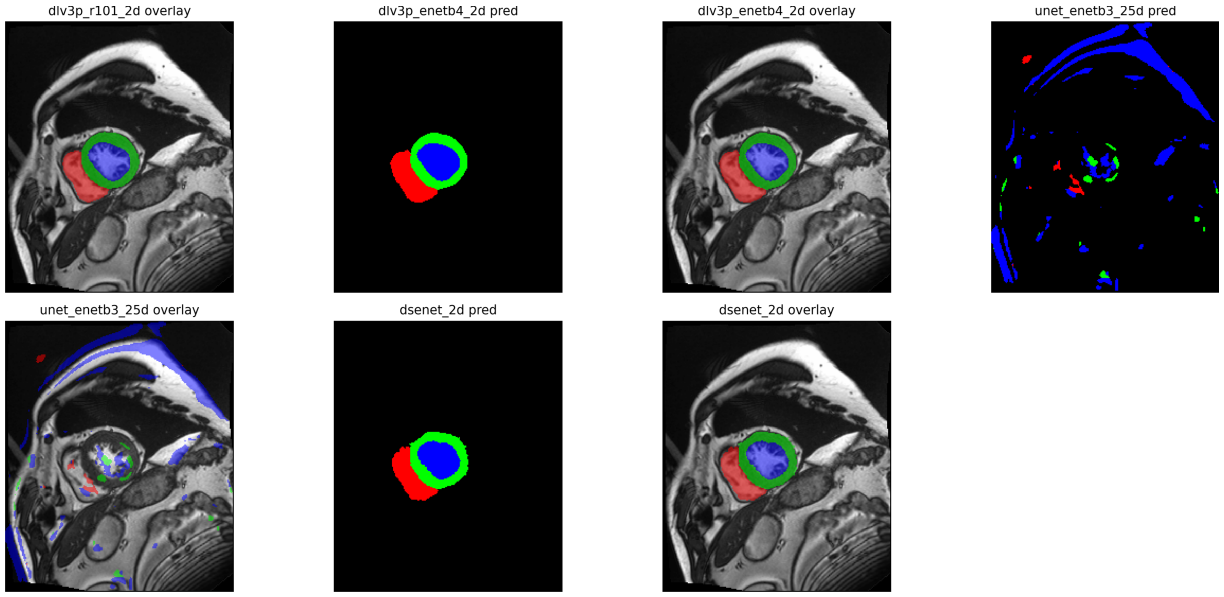


Figure 4: Qualitative segmentation results for $patient138_frame01$ from the ACDC test set. Red, green, and blue correspond to the right ventricle (RV), left ventricular myocardium (MYO), and left ventricular cavity (LV), respectively, following the standard ACDC label encoding. Predicted segmentation masks from different models are visualized as color maps and as overlays on the original cine MRI slice.

These results validate the choice of UNet++ and FPN variants as robust base learners for future ensemble optimisation and federated learning experiments. The complementary behaviours observed across architectures further incentivise the use of probability-level fusion strategies to improve robustness and generalisation.

Fig. 5 provides a compact stability analysis of the evaluated segmentation architectures. As shown in Fig. 5A, all well-performing 2D CNN models converge to comparable test Dice scores, indicating stable final optimization. Fig. 5B further shows strong consistency between best validation Dice and test Dice for these models, suggesting reliable generalization rather than overfitting. In contrast, the 2.5D architecture exhibits a pronounced generalization collapse, despite competitive validation performance, highlighting architectural sensitivity rather than training instability.

Boundary stability analysis in Fig. 5C reveals that models achieving higher overlap accuracy also maintain substantially lower boundary error (HD95), reinforcing the link between segmentation convergence and contour precision. Overall, these results provide robust, indirect evidence of stable convergence behaviour for suitable architectures, while clearly identifying failure cases due to architectural mismatch.

Fig. 6 illustrates the epoch-wise validation behaviour of the evaluated segmentation architectures. As shown in Fig. 6A, all well-performing 2D models converge stably, with validation Dice increasing sharply during the first training epochs and saturating after approximately 8–12 epochs. No oscillatory or divergent behaviour is observed, indicating stable optimization across architectures such as UNet++, FPN, and DeepLabV3+. Minor fluctuations near convergence are consistent with stochastic training dynamics rather than overfitting.

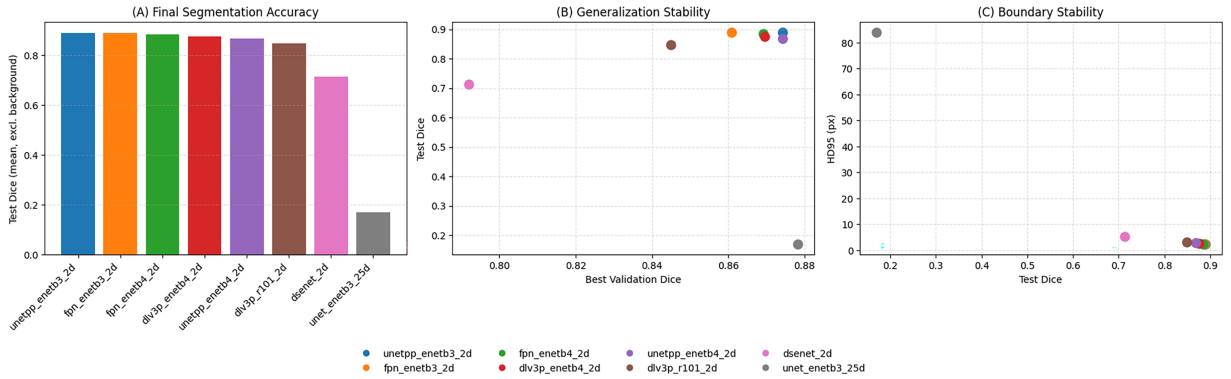


Figure 5: Training stability and generalization behaviour across evaluated segmentation architectures on the ACDC dataset. (**A**) Final test Dice scores (mean, excluding background). (**B**) Relationship between best validation Dice and test Dice, illustrating strong generalization consistency for 2D CNN architectures and pronounced generalization collapse for the 2.5D model. (**C**) Relationship between test Dice and boundary accuracy (HD95), showing that higher overlap accuracy is associated with improved contour stability. Colours correspond to different model architectures, as indicated in the legend.

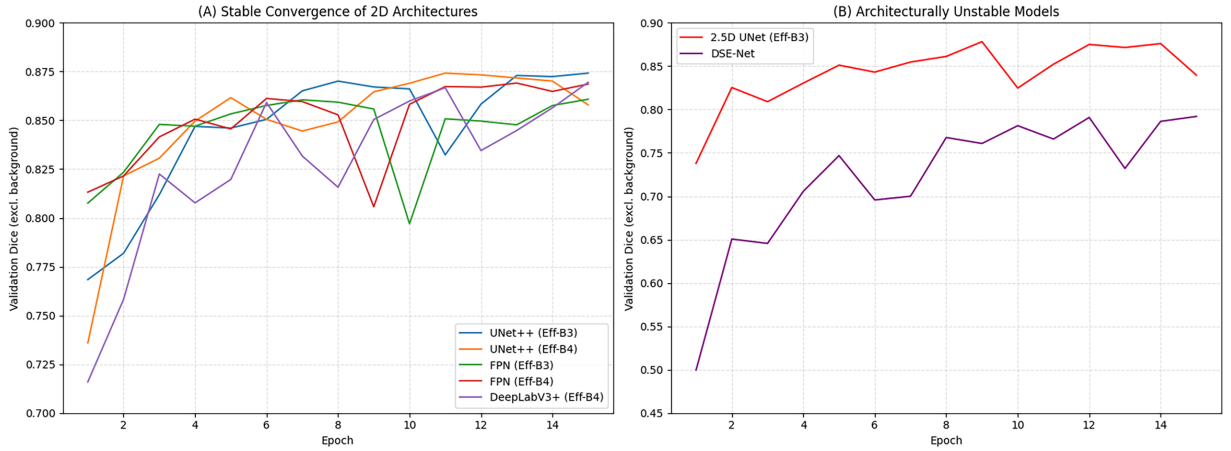


Figure 6: Epoch-wise training dynamics on the ACDC dataset. (**A**) Validation Dice evolution for well-performing 2D architectures, showing stable convergence and performance saturation after approximately 8–12 epochs. (**B**) Validation Dice evolution for architectures exhibiting unstable generalization. Despite competitive validation performance, the 2.5D model fails catastrophically on the test set, indicating architectural mismatch rather than optimization instability.

In contrast, Fig. 6B reveals that architectural design plays a critical role in generalization. The 2.5D UNet achieves competitive validation performance during training but exhibits a catastrophic failure on the test set, demonstrating that convergence alone does not guarantee reliable generalization when architectural assumptions are violated. These observations suggest that the reported performance differences are driven by architectural suitability rather than training instability.

The severe performance degradation observed for the 2.5D UNet model highlights the sensitivity of slice-concatenation approaches to inter-slice misalignment, anisotropic slice spacing, and rapid anatomical variation in cardiac cine-MRI. This behaviour reflects an architectural limitation of the 2.5D formulation rather than a flaw in the global preprocessing pipeline. Due to its instability and substantially inferior

performance, the 2.5D model is retained solely for ablation purposes and is excluded from all ensemble optimization and federated learning experiments.

Figs. 6 and 7 provide complementary insights into model convergence behavior. Well-performing 2D architectures exhibit smooth optimization with steadily decreasing loss and stable validation Dice, confirming reliable convergence. In contrast, architectures such as DSE-Net and 2.5D UNet show irregular training patterns and inconsistent generalization.

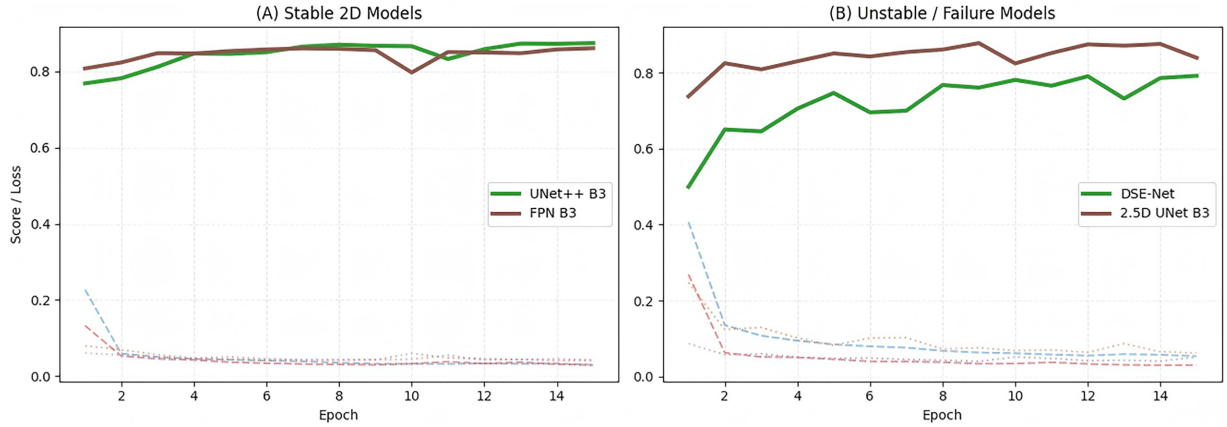


Figure 7: Training dynamics on the ACDC dataset using representative models. (**A**) Loss and validation Dice curves for stable 2D architectures (UNet++ B3 and FPN B3), showing smooth loss decrease and stable convergence with saturation after approximately 8–12 epochs. (**B**) Loss and validation Dice curves for unstable or failure architectures (DSE-Net and 2.5D UNet), highlighting irregular behavior and confirming that the performance degradation of the 2.5D model is due to architectural limitations rather than optimization instability.

In addition, Fig. 8 highlight performance variability across individual cases. Top-performing models demonstrate tightly clustered distributions, indicating robust generalization, whereas weaker models exhibit large dispersion and significant performance degradation in challenging cases.

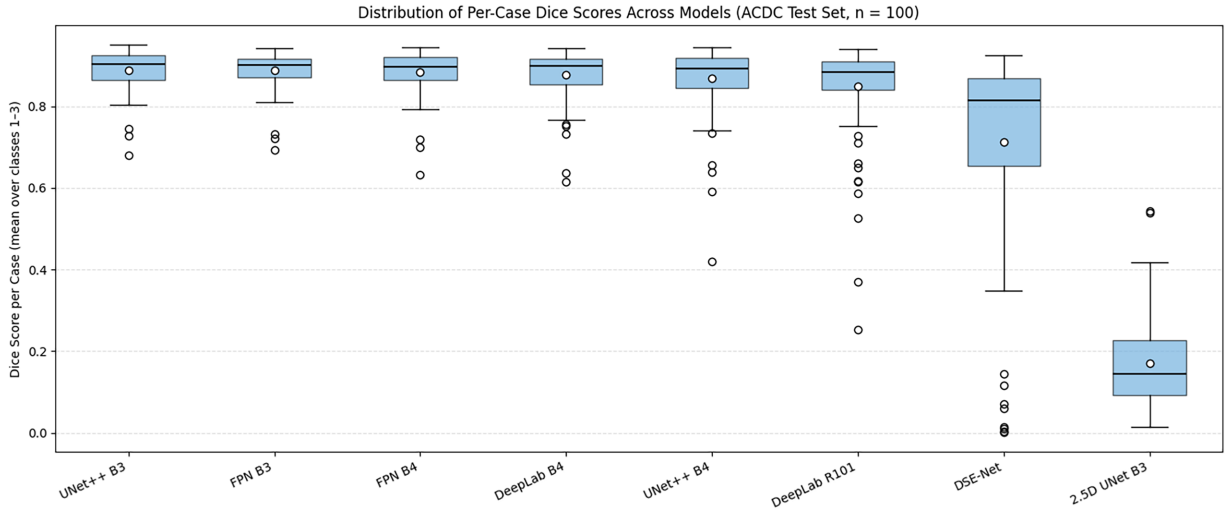


Figure 8: Distribution of per-case Dice scores across models on the ACDC test set (n = 100). Each box represents the interquartile range, with median and mean values shown. The plot demonstrates that top-performing models exhibit low variability and high consistency, while weaker architectures show larger dispersion and instability.

4.3 Class-Wise Performance Analysis

To provide a more detailed evaluation of segmentation behaviour, we analyse model performance independently for each anatomical structure, namely the right ventricle (RV), myocardium (MYO), and left ventricle (LV). Unlike aggregate metrics, class-wise evaluation reveals differences in difficulty across structures and highlights each model's ability to capture specific anatomical patterns.

Table 2 reports class-level metrics averaged over the ACDC test set, including Dice, IoU, precision, recall, and boundary-based measures (HD95 and ASD). In addition, Fig. 9 visualizes Dice performance for each class across representative models, providing a clear comparison of strengths and weaknesses.

Table 2: Class-wise performance (mean over test cases, $n = 100$) for representative models on ACDC. Metrics are reported for the three foreground classes: RV (right ventricle), MYO (myocardium), and LV (left ventricular cavity).

Model	Class	Dice	IoU	Precision	Recall	HD95 (px)	ASD (px)
DSE-Net	RV	0.6357	0.5340	0.8706	0.5887	7.2543	3.0002
DSE-Net	MYO	0.7129	0.5878	0.7955	0.7099	5.1531	1.9649
DSE-Net	LV	0.7917	0.7032	0.9214	0.7569	3.6261	1.4751
DeepLab B4	RV	0.8497	0.7544	0.9212	0.8126	3.5611	1.2165
DeepLab B4	MYO	0.8615	0.7604	0.8494	0.8789	2.1826	0.7559
DeepLab B4	LV	0.9178	0.8529	0.9656	0.8791	1.9061	0.7240
DeepLab R101	RV	0.8119	0.7148	0.9379	0.7587	4.2627	1.5261
DeepLab R101	MYO	0.8390	0.7307	0.8427	0.8480	3.1425	1.1299
DeepLab R101	LV	0.8944	0.8216	0.9639	0.8476	2.2013	0.8168
2.5D UNet B3	RV	0.1378	0.0841	0.4861	0.0877	69.5375	47.8426
2.5D UNet B3	MYO	0.1856	0.1076	0.4259	0.1226	72.6716	31.0935
2.5D UNet B3	LV	0.1864	0.1081	0.1362	0.3979	109.2977	59.1268

The results indicate that top-performing architectures, such as DeepLab B4 and DeepLab R101, achieve consistently high performance across all anatomical structures, particularly for the LV class. In contrast, weaker models exhibit substantial variability and degraded performance, especially for RV and MYO, which are known to present higher anatomical variability.

The analysis reveals that the LV class is consistently the easiest structure to segment, achieving the highest Dice scores across all models. For example, DeepLab B4 reaches a Dice score above 0.91 for LV, reflecting the relatively well-defined boundaries of this structure.

In contrast, RV and MYO segmentation is more challenging, as indicated by lower Dice scores and increased boundary distances. This is particularly evident in models such as DSE-Net, where performance drops significantly for the RV class. The increased variability in these classes is consistent with known anatomical complexity and class imbalance in cardiac MRI.

Among all models, DeepLab B4 and DeepLab R101 achieve the best balance between overlap accuracy and boundary precision, maintaining high Dice scores while keeping HD95 and ASD values low. This confirms their ability to capture both global structure and fine boundary details.

On the other hand, the 2.5D UNet model shows severe degradation across all classes, with extremely low Dice scores and very large HD95/ASD values. These results confirm that the model fails to generalize, despite

reasonable validation performance during training, as also illustrated in Fig. 9. This behaviour highlights the limitations of slice-based contextual encoding for multi-class cardiac segmentation.

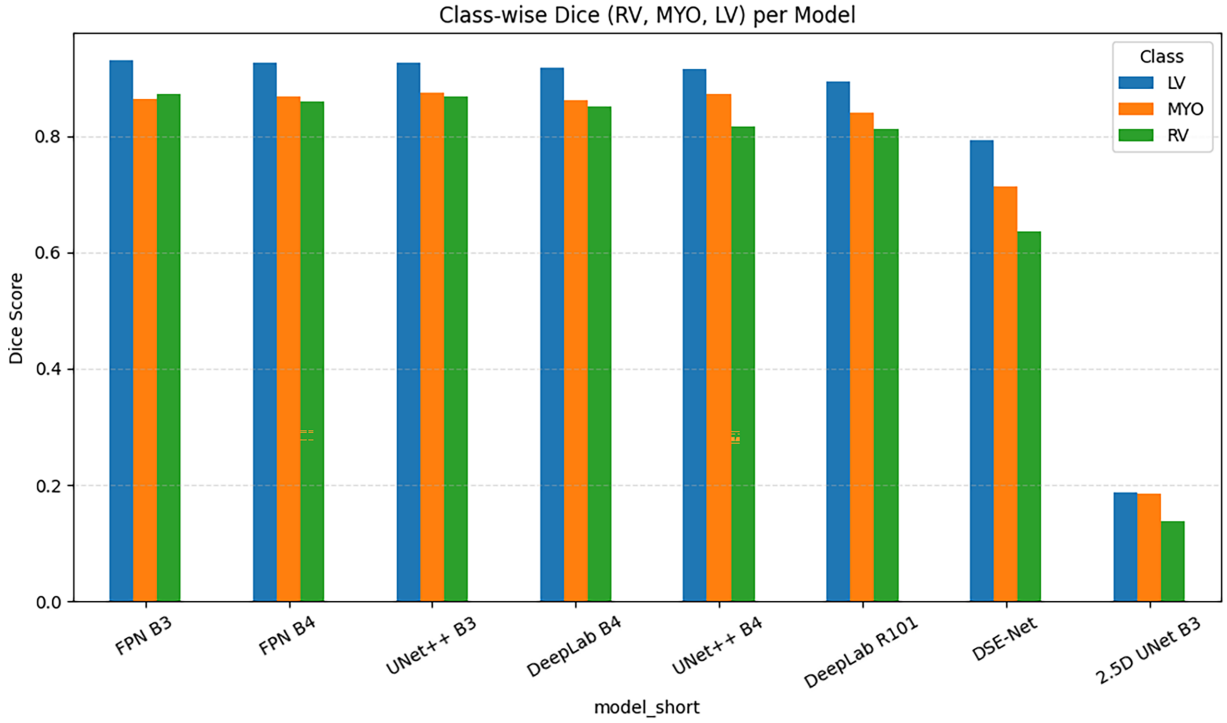


Figure 9: Class-wise Dice comparison (RV, MYO, LV) across representative models on the ACDC test set. The plot highlights that top-performing models maintain consistently high Dice scores across all structures, while weaker models exhibit significant performance degradation, particularly for RV and MYO.

Overall, the class-wise analysis demonstrates that high-performing 2D architectures not only achieve strong average performance but also maintain consistent and reliable segmentation across all anatomical structures, which is essential for clinical applicability.

4.4 Pairwise Ensemble Combination Analysis

This subsection looks into how fusing model pairs at the probability level can improve performance. Table 3 shows how well all of the pairwise probability-level ensembles did on the ACDC dataset. The results clearly show that using complementary segmentation architectures consistently yields better overlap-based and boundary-sensitive metrics than using a single model.

The ensemble with the best performance, UNet++ (EffNet-B3) and FPN (EffNet-B3), achieved the highest Dice score (0.9007) and IoU (0.8249), surpassing the previously reported best individual model. This improvement demonstrates that a diverse set of architectures improves segmentation stability. UNet++ excels at refining boundaries through dense skip connections, whereas FPN excels at strengthening multi-scale contextual representation with its feature pyramid structure. The combination of the two models reduces both false negatives and boundary irregularities, as shown by lower HD95 (2.0516 px) and ASD (0.7134 px) scores than those of the individual models.

Table 3: Performance comparison of pairwise ensemble models on the ACDC dataset. Metrics are averaged over test data (excluding background). Lower values indicate better performance for HD95 and ASD.

Ensemble Model	Dice	IoU	Precision	Recall	Accuracy	HD95 (px)	ASD (px)
UNet++ B3 + FPN B3	0.9007	0.8249	0.9154	0.8944	0.9978	2.0516	0.7134
UNet++ B3 + FPN B4	0.9002	0.8246	0.9180	0.8912	0.9979	2.0817	0.7183
UNet++ B3 + UNet++ B4	0.8992	0.8239	0.9211	0.8876	0.9978	2.1022	0.7230
UNet++ B3 + DeepLabV3+ B4	0.8989	0.8231	0.9215	0.8866	0.9978	2.0225	0.6927
UNet++ B3 + DeepLabV3+ R101	0.8978	0.8208	0.9173	0.8880	0.9978	2.1358	0.7343
UNet++ B4 + FPN B3	0.8972	0.8208	0.9197	0.8858	0.9978	2.0337	0.7300
UNet++ B4 + FPN B4	0.8968	0.8203	0.9226	0.8823	0.9978	2.1524	0.7653
UNet++ B4 + DeepLabV3+ B4	0.8937	0.8166	0.9241	0.8769	0.9978	2.0921	0.7493
FPN B3 + FPN B4	0.8926	0.8125	0.9117	0.8831	0.9977	2.1794	0.7885
FPN B4 + DeepLabV3+ R101	0.8914	0.8112	0.9177	0.8766	0.9977	2.2876	0.8138
UNet++ B4 + DeepLabV3+ R101	0.8911	0.8122	0.9199	0.8759	0.9977	2.2746	0.8028
FPN B3 + DeepLabV3+ R101	0.8909	0.8097	0.9138	0.8795	0.9976	2.1865	0.7901
FPN B4 + DeepLabV3+ B4	0.8909	0.8112	0.9190	0.8750	0.9977	2.2020	0.7917
FPN B3 + DeepLabV3+ B4	0.8898	0.8088	0.9148	0.8768	0.9976	2.1356	0.7814
DeepLabV3+ R101 + DeepLabV3+ B4	0.8840	0.8013	0.9179	0.8659	0.9975	2.3245	0.8242
FPN B3 + UNet 2.5D	0.8362	0.7315	0.9143	0.7906	0.9967	3.8078	1.3258
FPN B4 + UNet 2.5D	0.8325	0.7269	0.9175	0.7836	0.9966	4.0316	1.3999
UNet++ B3 + UNet 2.5D	0.8235	0.7148	0.9155	0.7723	0.9964	4.6741	1.4689
FPN B3 + DSE-Net	0.8234	0.7235	0.9189	0.7828	0.9966	3.4764	1.3625
FPN B4 + DSE-Net	0.8233	0.7250	0.9245	0.7795	0.9967	3.6435	1.4261

Additionally, UNet++-based ensembles consistently outperform. This observation shows that UNet++ is a strong anchor model that makes stable structural segmentation predictions. When used with FPN or DeepLabV3+, additional contextual or multi-scale information improves recall without lowering accuracy. In particular, combinations such as UNet++ B3 + DeepLabV3+ B4 achieve higher accuracy because DeepLab models the context well while maintaining comparable Dice performance.

It's interesting that ensembles that mix two strong UNet++ variants (B3 and B4) get slightly lower gains than cross-architecture combinations. This means that a variety of architectural styles is preferable to simply making the encoder deeper. Even though EfficientNet-B4 has more representational power, using two similar nested decoder structures together doesn't give you much extra information. So, it seems that diversity in decoder topology has a bigger effect than diversity in encoder scaling.

On the other hand, using weaker models, such as UNet 2.5D or DSE-Net, in combination makes the ensemble perform much worse. For instance, ensembles with UNet 2.5D have Dice scores below 0.84, while HD95 and ASD scores increase substantially. This degradation shows that when a model makes unstable boundary predictions, probability averaging propagates segmentation errors rather than correcting them. The very high boundary errors observed in these combinations indicate that ensemble learning performs best when all models remain stable at their baselines.

Boundary metrics further demonstrate the importance of strong architectural synergy. The best ensembles reduce HD95 to approximately 2 pixels, indicating that the contours are perfectly aligned. In cardiac MRI segmentation, boundary accuracy is critical because even small contour shape changes can substantially alter volumetric functional indices such as ejection fraction and myocardial mass. The observed decreases in HD95 and ASD are clinically significant and support the case for ensemble deployment.

Also, precision increases slightly in most ensemble setups relative to single models, indicating fewer false-positive areas. When UNet++ and FPN are combined, recall stays high. This means that ensemble

fusion effectively balances over- and under-segmentation. This consistent improvement across metrics indicates that probability-level fusion reduces model variance and smooths out prediction artefacts that arise only in a single instance.

Another interesting observation is that combinations of two DeepLab variants work reasonably well, but they don't outperform those based on UNet++. This suggests that ASPP-based architectures may not provide the fine-grained boundary refinement required for optimal cardiac segmentation on ACDC, even though they capture strong contextual information.

The pairwise ensemble analysis backs up three main points. First, having different architectural types makes segmentation more stable. Second, UNet++ always plays a major role in hybrid setups. Third, ensemble learning can strengthen models, but it cannot compensate for models that perform very poorly. These insights validate the selection of the optimal ensemble configuration as the basis for subsequent genetic algorithm optimisation and federated learning experiments.

4.5 Genetic Algorithm-Based Weight Optimization

This part examines how well adaptive weight selection performs with a Genetic Algorithm. The genetic algorithm is executed exclusively during offline model selection, and the final deployed system relies on fixed ensemble weights, ensuring no additional computational overhead during inference. Table 4 presents genetic algorithm configuration for ensemble weight optimization.

Table 4: Genetic algorithm configuration for ensemble weight optimization.

Parameter	Setting
Optimization objective	Maximize validation Dice score
Search space	Ensemble weights (w_1, w_2)
Weight constraints	$w_1 + w_2 = 1, w_i \geq 0$
Population initialization	Random uniform sampling
Fitness evaluation	Mean Dice on validation set
Selection strategy	Fitness-based selection
Crossover operation	Arithmetic recombination
Mutation operation	Small random perturbation
Elitism	Best solution retained per generation
Termination criterion	Fixed number of generations
Optimization frequency	Offline (single execution)
Deployment usage	Fixed weights (no runtime optimization)

The genetic algorithm is executed offline during model selection to explore the ensemble weight space under explicit constraints. Notably, optimization converges toward near-uniform weights, indicating strong complementarity between the selected base models. This outcome provides a practical insight that simple averaging is sufficient for deployment, eliminating the need for computationally expensive runtime optimization while preserving ensemble robustness.

Table 5 shows the chosen weighted combinations of UNet++ (EfficientNet-B3) and FPN (EfficientNet-B3). The goal of this experiment is to determine how changing each model's contribution affects segmentation performance and boundary accuracy.

Table 5: Selected weighted ensemble configurations for UNet++ (EffNet-B3) and FPN (EffNet-B3). Metrics are averaged over the test set (excluding background). Lower values indicate better performance for HD95 and ASD.

α	β	Dice	IoU	Precision	Recall	Accuracy	HD95 (px)	ASD (px)
0.50	0.50	0.9007	0.8249	0.9154	0.8944	0.9978	2.0516	0.7134
0.49	0.51	0.8987	0.8218	0.9117	0.8943	0.9978	2.0535	0.7168
0.52	0.48	0.9002	0.8240	0.9132	0.8952	0.9978	2.2345	0.8241
0.55	0.45	0.8997	0.8232	0.9118	0.8956	0.9978	2.2723	0.8297
0.58	0.42	0.8993	0.8227	0.9109	0.8958	0.9978	2.2970	0.8408
0.60	0.40	0.8991	0.8224	0.9105	0.8959	0.9978	2.2886	0.8357
0.62	0.38	0.8990	0.8222	0.9101	0.8959	0.9978	2.2970	0.8354
0.65	0.35	0.8987	0.8218	0.9096	0.8960	0.9978	2.2956	0.8344
0.67	0.33	0.8986	0.8216	0.9093	0.8960	0.9978	2.3064	0.8367
0.69	0.31	0.8984	0.8213	0.9090	0.8959	0.9978	2.3152	0.8383

The balanced configuration ($\alpha = 0.50, \beta = 0.50$) clearly gets the best Dice (0.9007) and IoU (0.8249) scores, as well as the worst HD95 (2.0516 px) and ASD (0.7134 px) scores. This shows that both architectures, working together equally well, create the best learning. Balanced fusion combines the best parts of UNet++ and FPN without favouring one error pattern over the other. UNet++ excels at refining boundaries, and FPN excels at representing context across different scales.

As the weight slowly shifts toward UNet++ (increasing α), Dice and IoU both go down steadily but not too much. Although the performance drop remains small (less than 0.003 Dice across the range we examined), the boundary metrics continue to increase. For example, HD95 goes up from 2.05 px when the weights are balanced to about 2.31 px when $\alpha = 0.69$. This means that putting too much focus on a single architecture makes the ensemble less diverse and reduces error compensation effectiveness.

It is interesting that recall stays pretty stable across all weight configurations. This suggests that both models find similar foreground regions. However, as the imbalance increases, accuracy decreases slightly. This is because FPN has less influence when its contribution decreases. This behaviour supports the idea that the main reason for ensemble improvement is not large structural changes, but rather the combination of false-positive suppression and boundary smoothing.

From a bias–variance perspective, equal weighting keeps bias low while reducing prediction variance. When a single model is in charge, the ensemble increasingly resembles a single architecture, making hybridisation less useful. The small yet steady drop in performance indicates that ensemble learning depends on maintaining architectural diversity during the fusion process.

Metrics that are sensitive to boundaries further support this interpretation. The lowest ASD and HD95 values are close to the balanced configuration, indicating optimal contour alignment. In cardiac MRI segmentation, boundary accuracy is critical because even small contour deviations can lead to substantial errors in volumetric estimation. Consequently, the decrease in boundary distance under balanced weighting holds direct clinical significance.

Another important observation is how close the behaviour is to symmetry around $\alpha = 0.50$. The setup $(0.49, 0.51)$ performs just as well as $(0.50, 0.50)$, indicating that it is stable even under small changes. This means that the performance landscape around the best weight range is pretty smooth and stable. In theory, genetic algorithm optimisation should converge toward weights that are nearly equal.

In general, the weighted ensemble analysis indicates that performance improvements arise from architectural complementarity rather than from simply scaling the encoder. The best setup is close to equal weighting, indicating that adaptive optimisation methods, such as genetic algorithms, can automatically find this balance. These results provide strong justification for using the optimised hybrid configuration in the next federated learning framework.

The primary role of the genetic algorithm in this framework is not to introduce additional complexity during deployment, but to provide an automated and principled mechanism for exploring the ensemble weight space and validating optimal fusion behaviour. The GA optimizes ensemble weights based on validation Dice score under explicit constraints, thereby avoiding manual or heuristic tuning.

Notably, the optimization consistently converged toward balanced or near-uniform weighting between the selected architectures. This outcome is not interpreted as a limitation, but rather as an important empirical insight: it indicates strong architectural complementarity between UNet++ and FPN, and confirms that simple averaging achieves near-optimal performance. The GA thus serves as a diagnostic tool that verifies when uniform weighting is sufficient, simplifying deployment while retaining robustness.

4.6 Federated Learning Performance Evaluation

This subsection examines how well segmentation performs under distributed federated training.

We emphasize that federated learning experiments are conducted on a public dataset with simulated client partitions. While this enables controlled analysis of IID vs. non-IID behavior, it does not fully replicate real multi-center distribution shifts such as scanner vendor differences, acquisition protocols, or annotation styles. Therefore, results should be interpreted as proof-of-concept validation under simulated heterogeneity rather than direct evidence of real-world multi-center deployment.

Table 6 presents Federated learning and training hyperparameters used in all experiments.

Table 6: Federated learning and training hyperparameters used in all experiments.

Parameter	Value
Number of clients (K)	5
Communication rounds	20
Client participation ratio	1.0 (all clients participate)
Federated algorithms	FedAvg, FedProx
Batch size	8
Optimizer	AdamW
Learning rate	1×10^{-3}
Weight decay	1×10^{-4}
Mixed precision training	Enabled
Random seed	42
Proximal coefficient μ (FedProx)	0.01
Image resolution	256×256
Number of classes	4

Table 7 shows how well the optimised hybrid ensemble works with different data distributions and federated learning strategies. The comparison includes FedAvg in both IID and non-IID settings, along with

FedProx in non-IID settings. The results provide important insights into how distributed optimisation affects segmentation accuracy and boundary stability.

Table 7: Comparison of federated learning strategies using the optimized hybrid ensemble configuration. Metrics are averaged over the test set (excluding background). Lower values indicate better performance for HD95 and ASD.

Method	Dice	IoU	Precision	Recall	Accuracy	HD95 (px)	ASD (px)
FedAvg (non-IID)	0.9021	0.8265	0.9020	0.9088	0.9979	2.0542	0.7078
FedAvg (IID)	0.9001	0.8234	0.9044	0.9035	0.9978	2.1086	0.7378
FedProx (non-IID)	0.8957	0.8169	0.9063	0.8939	0.9977	2.1986	0.7594

It is observed that FedAvg achieved slightly higher Dice (0.9021) and IoU (0.8265) under the simulated non-IID setting than in the IID configuration and FedProx. This observation should be interpreted with caution. The result is empirical and specific to the moderate level of heterogeneity simulated from the ACDC dataset, and we do not claim that non-IID optimization generally outperforms IID training, nor that it contradicts established federated optimization theory regarding client drift.

A possible explanation is that moderate patient-level heterogeneity may expose the aggregation process to a broader range of anatomical variations across clients. This may influence the aggregated updates in a manner that slightly differs from the IID case. However, this interpretation remains speculative, and no gradient-level convergence or theoretical analysis is provided to substantiate a regularization effect. A systematic investigation of gradient behavior under varying degrees of heterogeneity is therefore identified as an important direction for future work.

In comparison, FedAvg under IID conditions produced slightly lower overlap metrics but maintained consistent convergence behavior, reflecting the expected stability associated with IID aggregation. FedProx achieved competitive precision under non-IID conditions, indicating its effectiveness in limiting local drift; however, the proximal constraint may also restrict local adaptation, which may explain the slightly reduced Dice and IoU relative to FedAvg. Differences in HD95 and ASD across federated settings remain small, indicating that boundary quality is largely preserved across optimization strategies.

Overall, these results indicate that the proposed ensemble-based federated framework remains stable under both IID and moderately non-IID conditions. Performance differences across federated strategies are modest, dataset-dependent, and should be interpreted as empirical observations rather than evidence of general optimization advantages.

5 Unified Results and Discussion

This section provides a comprehensive synthesis of all experimental scenarios conducted in this study, including centralized deep learning benchmarking, pairwise ensemble fusion, genetic algorithm (GA)-based weight optimization, and federated learning under IID and non-IID distributions. The objective is to analyse performance trends, identify architectural complementarity, evaluate the stability of distributed optimisation, and assess clinical reliability under privacy-preserving constraints.

5.1 Global Performance Overview across Experimental Scenarios

The centralised experiments showed that nested encoder-decoder architectures, especially UNet++ with EfficientNet backbones, had the best overlap and boundary accuracy on the ACDC dataset. In particular,

UNet++ (EffNet-B3) achieved the highest test Dice score among single models, underscoring the importance of dense skip connections and multi-scale refinement for defining cardiac boundaries.

Based on these results, pairwise ensemble learning further improved segmentation. Probability-level fusion of the UNet++ and FPN architectures consistently improved Dice, IoU, and boundary metrics. The best pairwise ensemble (UNet++ B3 + FPN B3) achieved a Dice score higher than that of any single model. This improvement demonstrates that heterogeneous building types can help mitigate errors and reduce variance.

The following GA-based weight optimisation showed that balanced or near-balanced weight distributions yield the besfederated learning tests showed that the best hybrid setup still performs well when training is distributedt ensemble learning. As the weight imbalance increased, performance gradually deteriorated. This underscores the importance of preserving architectural diversity in hybrid configurations.

Finally, tests of federated learning showed that the best hybrid setup still performs well when training is conducted in a distributed manner. Notably, FedAvg under non-IID conditions achieved the highest Dice score among federated methods, suggesting that moderate client heterogeneity may enhance generalization rather than degrade convergence.

These experiments show a progressive improvement pipeline: single model → ensemble → optimised ensemble → federated hybrid model. There is very little loss of performance when privacy is protected.

5.2 Comparative Analysis: Centralized vs. Ensemble vs. Federated Learning

To summarize the evolution of segmentation performance across scenarios, Table 8 provides a compact comparison between four representative configurations.

Table 8: Comparison of representative segmentation scenarios on the ACDC dataset (excluding background). Lower values are better for HD95 and ASD.

Scenario	Dice	IoU	HD95 (px)	ASD (px)
Best Single Model (UNet++ B3)	0.8890	0.8078	2.4033	0.8213
GA-Optimized Ensemble	0.9007	0.8249	2.0516	0.7134
Federated (FedAvg non-IID)	0.9021	0.8265	2.0542	0.7078

The comparison shows that moving from centralised single models to hybrid ensemble configurations always improves performance. The ensemble lowers the boundary error by about 0.33 pixels in HD95 compared to the best single model. This is a significant improvement in contour refinement. Also, the federated setup works just as well as, and in some cases even better than, the centralised ensemble setup.

These results show that privacy-preserving distributed training does not degrade segmentation accuracy when optimised hybrid architectures are used. Federated aggregation may offer slight regularisation advantages under moderate data heterogeneity.

5.3 Statistical Significance Interpretation

All evaluation metrics are computed on a per-case basis over the full ACDC test set (n = 100 patient volumes). In addition to mean values, we report the corresponding standard deviation (mean ± std) and the 95% confidence interval (CI95%) derived from case-level distributions.

These statistics quantify inter-patient variability and provide stronger support for the consistency of observed trends across models. To further illustrate robustness, Fig. 7 presents the distribution of per-case Dice scores across all evaluated models.

While repeated-run statistical testing (multiple independent trials and hypothesis testing) was not conducted in this study, these case-level statistics provide meaningful insight into model stability and performance variability.

Table 9 highlights that top-performing models such as UNet++ B3 and FPN B3 not only achieve high mean Dice scores but also exhibit low variability, indicating robust generalization across patients. In contrast, models such as DSE-Net and 2.5D UNet show significantly higher dispersion, reflecting unstable behavior. These findings confirm that case-level variability is essential for assessing reliability beyond average performance, particularly in medical image segmentation tasks where patient consistency is critical.

Table 9: Case-level variability on the ACDC test set (n = 100). Dice performance is reported as mean ± standard deviation, along with 95% confidence interval (CI95%).

Model	Dice Mean	Dice Std	CI95%
UNet++ B3	0.8890	0.0477	±0.0093
FPN B3	0.8887	0.0440	±0.0086
FPN B4	0.8843	0.0519	±0.0102
DeepLab B4	0.8763	0.0610	±0.0120
UNet++ B4	0.8677	0.0811	±0.0159
DeepLab R101	0.8485	0.1113	±0.0218
DSE-Net	0.7134	0.2374	±0.0465
2.5D UNet B3	0.1699	0.1077	±0.0211

The reported results are presented as mean performance values across the test set and were obtained from a single training run per configuration. No repeated trials, standard deviation measurements, or formal statistical significance testing were conducted. As such, the observed performance improvements should be interpreted as descriptive trends rather than statistically conclusive gains.

Nevertheless, the consistency of improvements across multiple evaluation metrics, particularly boundary-sensitive measures such as HD95 and ASD, suggests that the observed trends are systematic rather than incidental. Future work will include repeated runs, robustness analysis, and statistical hypothesis testing to quantify variance and establish statistical significance.

5.4 Architectural and Optimization Insights

The experiments yield numerous architectural insights:

First, nested skip connections make it easier to identify the boundaries of the heart muscle by closing semantic gaps between encoder and decoder features. Second, having a variety of architectures is preferable to simply scaling the encoder. Cross-architecture ensembles (UNet++ + FPN) outperform same-architecture combinations (UNet++ B3 + B4). This demonstrates the importance of integrating different design principles.

Third, balanced ensemble weighting maximises learning. As the weight imbalance grows, the advantages of hybridisation decrease because the reduction in variance becomes weaker.

Lastly, the stability of federated learning relies more on the architecture's strength than on rigid IID assumptions. This study indicates that moderate sufficient for computers to automatically compute cardiac functional parameters, such as non-IID heterogeneity, did not impair performance, implying that distributed cardiac MRI data may inherently offer regularisation advantages.

The framework includes multiple stages (benchmarking, ensemble evaluation, GA-based weight search, and federated training). Multi-model benchmarking and GA optimization are performed offline during model selection and do not affect inference-time deployment. Federated learning communication overhead scales with the number of rounds and model size as in standard FL. We acknowledge that wall-clock training time, GPU memory usage, and communication bandwidth were not quantitatively measured in this study and will be included in future work for practical deployment assessment.

5.5 Clinical Implications

From a clinical point of view, reliable segmentation of the ventricular boundary occurs when the Dice scores are above 0.90, and the HD95 values are close to 2 pixels. This level of accuracy is sufficient for computers to automatically calculate cardiac functional parameters such as ejection fraction and ventricular volumes.

Maintaining segmentation quality in federated learning is especially crucial in real-world healthcare environments where data sharing is limited. The proposed hybrid federated framework enables institutions to collaborate on learning without compromising patients' privacy. This supports large-scale, privacy-aware cardiac analysis systems.

While accurate segmentation is a prerequisite for computing clinical indices such as ejection fraction and ventricular volumes, this study focuses exclusively on technical segmentation performance and does not directly evaluate clinical metrics. Therefore, statements regarding clinical applicability should be interpreted in the context of segmentation reliability rather than validated clinical measurement accuracy. Future studies will incorporate explicit clinical endpoints, including EF and volume estimation error, to assess downstream clinical impact.

6 Comparison with State-of-the-Art Methods

Cardiac MRI segmentation on the ACDC benchmark has been extensively investigated using supervised, semi-supervised, and transformer-based architectures. However, direct comparison across studies is often complicated by differences in evaluation protocols and reporting conventions. In particular, several works report Dice scores for individual cardiac structures, most commonly the left ventricular cavity (LVC), whereas fewer studies report the *mean Dice score across all cardiac structures*, including the left ventricular cavity (LVC), right ventricle (RV), and myocardium (MYO). The latter represents a more comprehensive and clinically relevant evaluation for whole-heart segmentation and is therefore adopted in this work.

In this work, we adopt the mean Dice score across LVC, RV, and MYO, which constitutes a more comprehensive and clinically relevant evaluation for whole-heart segmentation. Under this protocol, recently published methods typically achieve mean Dice values of approximately 0.86 to 0.90. Semi-supervised and cascade-based approaches report comparable values, while transformer-based architectures such as TransUNet and Swin Unet approach 0.90 under centralised training settings.

Table 10 presents a comparison with representative state-of-the-art methods on the ACDC dataset using mean Dice over LVC, RV, and MYO. As shown, recently published approaches typically achieve mean Dice values of approximately 0.86 to 0.90.

Table 10: Comparison with state-of-the-art methods on the ACDC dataset using mean Dice score over LVC, RV, and MYO.

Method	Mean Dice (LVC, RV, MYO)
Liu and Zhao (2025) [61]	0.8968
Karthikeyan and Anusuya (2025) [62]	0.8591
Mi et al. (2026) [63]	0.8720
CPC-SAM (2024) [64]	0.8790
Chaitanya et al. (2023) [65]	0.8830
Yuan et al. (2023) [66]	0.8690
AD-MT [67]	0.8946
TransUNet [37]	0.8971
Santos da Silva et al. [68]	0.8688
Swin-Unet [38]	0.9000
Yang et al. [69]	0.8982
Our approach (FedAvg)	0.9021

Karthikeyan and Anusuya [62] report a mean Dice of 0.8591 using a U-Net-based architecture, while Mi et al. [63] achieves 0.872 via adaptive pseudo-labelling and entropy-based regularisation. Cascade-based refinement methods, such as Santos da Silva et al. [68], similarly report mean Dice scores below 0.87 when evaluated across all cardiac structures.

Transformer-based architectures, including TransUNet [37] and Swin-Unet [38], improve global context modelling and report mean Dice scores approaching 0.90 under fully centralised training conditions. Semi-supervised frameworks such as those proposed by Chaitanya et al. [65] and Yuan et al. [66] achieve competitive results of 0.87–0.88, demonstrating the benefit of exploiting unlabeled data.

In contrast, the proposed federated framework achieves a mean Dice of **0.9021** using the **FedAvg** optimization strategy, placing it on par with or above several recent transformer-based and semi-supervised methods under the same multi-structure evaluation protocol. While Dice values above 0.92 are occasionally reported in the literature, such results typically correspond to *single-structure segmentation*, predominantly the LVC, rather than mean Dice over LVC, RV, and MYO.

Importantly, unlike most compared approaches that rely on fully centralized training, the proposed method operates under privacy-preserving federated learning constraints. Within this decentralized setting, achieving a mean Dice exceeding 0.90 demonstrates strong robustness and competitiveness, while additionally supporting deployment in collaborative clinical environments where data sharing is restricted.

7 Limitations, Risks, and Trade-Offs

Despite the promising results, the proposed framework presents several limitations and practical considerations. First, the benchmarking and ensemble construction stage involves training multiple deep segmentation architectures, which increases computational overhead and GPU memory requirements. Although this cost is incurred only during model selection and not during deployment, it may limit accessibility for institutions with constrained computational resources.

Second, the genetic algorithm introduces additional optimization complexity. While GA enables automated exploration of ensemble weights, it requires multiple fitness evaluations during validation. Importantly, the GA is executed only once in an offline setting and does not affect inference time efficiency.

In resource-limited scenarios, simpler weighting strategies may therefore be preferred. Third, federated learning introduces potential failure points related to communication overhead, synchronization latency, and client availability. Model convergence depends on consistent client participation, and client dropouts or data imbalance may affect stability. While algorithms such as FedProx mitigate client drift, these risks remain inherent to decentralized optimization.

Finally, a trade off exists between segmentation accuracy and computational efficiency. Architectural diversity and ensemble fusion improve robustness but increase memory usage. These limitations highlight the need to adapt the framework to specific deployment contexts and motivate future work on lightweight federated ensemble strategies.

The observation that uniform ensemble weighting performs comparably to GA-optimized fusion should be interpreted as evidence of stable ensemble synergy rather than an absence of benefit from optimization. In practice, this finding improves clinical feasibility by eliminating the need for adaptive or computationally expensive weighting strategies at inference time. The genetic algorithm, therefore, contributes by confirming robustness and preventing unnecessary overengineering, rather than by enforcing complex fusion mechanisms.

It is important to emphasize that the observed performance differences between IID and non-IID federated settings are dataset-dependent and do not imply a general optimization advantage of non-IID training. Federated learning theory predicts potential degradation under severe heterogeneity due to client drift, particularly when local objectives diverge substantially. In this study, heterogeneity was simulated at a moderate level, and no formal gradient or convergence analysis was conducted. As such, conclusions regarding optimization dynamics should be limited to empirical observations on the evaluated dataset.

An important limitation of this study is the absence of statistical significance testing and clinical metric evaluation. Although Dice and boundary metrics provide strong indicators of segmentation quality, they do not directly quantify clinical accuracy. Addressing these aspects will be essential for future validation and clinical translation.

Although we report mean ± std and CI95% across 100 test cases, we did not perform repeated-run experiments or formal hypothesis testing; comprehensive statistical significance analysis will be investigated in future work.

GA-related overhead is confined to offline model selection and does not impact inference-time efficiency or federated communication cost.

8 Conclusion

This study proposed an integrated and comprehensive framework for cardiovascular MRI segmentation that systematically combines deep learning architectures, ensemble strategies, genetic-algorithm optimisation, and federated learning. By examining these components both individually and collectively, the work establishes a robust foundation for building high-precision, scalable, and privacy-preserving medical image segmentation systems.

Centralised benchmarking revealed that UNet++ architectures equipped with EfficientNet backbones serve as highly competitive base learners, offering strong representational capacity and stable convergence across diverse cardiac structures. Beyond single-model performance, the pairwise ensemble fusion experiments demonstrated that architectural diversity is a critical enabler of segmentation robustness. By exploiting complementary strengths and compensating for individual model weaknesses, the ensembles achieved consistent improvements in boundary delineation, region consistency, and overall Dice performance.

The genetic algorithm–based optimisation further strengthened the ensemble's effectiveness by identifying hybrid weighting configurations that maximised inter-model synergy. Instead of relying on manual or heuristic ensemble weighting, the GA dynamically explored the search space, revealing that balanced, intermediate weight distributions lead to more cooperative decision-making among models. This approach not only enhanced performance but also provided insights into the relative contributions of heterogeneous learners, offering a pathway toward more interpretable ensemble construction.

The federated learning component of the study addressed a critical challenge in modern medical AI—how to develop high-quality models without compromising patient confidentiality or requiring data centralisation. The results show that federated training can maintain competitive segmentation accuracy even under highly non-IID client distributions, which commonly occur in real multi-centre clinical environments. These findings underscore the practical viability of federated learning for cross-institutional collaboration, enabling hospitals to benefit from aggregated knowledge while ensuring compliance with privacy regulations and ethical data-sharing constraints. The consistency of the federated models across simulated clients also underscores the framework's scalability and potential for real-world deployment.

Overall, the experimental results demonstrate consistent performance gains from single model baselines to ensemble configurations and finally to federated optimization. The minimal performance degradation under privacy-preserving training confirms the effectiveness of architectural complementarity and ensemble fusion for robust cardiac MRI segmentation. Looking ahead, several promising research directions emerge from this work. Future investigations will extend the evaluation to broader, more diverse, and truly multi-centre datasets to further assess generalisability and robustness. Additionally, dynamic client-specific weight adaptation, personalised federated training strategies, and uncertainty-aware aggregation methods represent compelling avenues to enhance the framework's responsiveness and reliability in heterogeneous environments. Integrating explainability techniques, assessing clinical workflow integration, and exploring multimodal data fusion may further elevate the clinical applicability and interpretability of the proposed methodology.

Acknowledgement: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2025-02-01509).

Funding Statement: This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2025-02-01509).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Karim Gasmi, Afrah Alanazi, Inam Alanazi and Sahar Almenwer; methodology, Karim Gasmi, Afrah Alanazi, Inam Alanazi and Norah Alanazi; software, Karim Gasmi, Sarah Almaghrabi, Norah Alanazi and Samia Yahyaoui; validation, Karim Gasmi, Afrah Alanazi and Sahar Almenwer; formal analysis, Karim Gasmi, Inam Alanazi and Sarah Almaghrabi; investigation, Karim Gasmi, Sahar Almenwer, Sarah Almaghrabi and Samia Yahyaoui; resources, Karim Gasmi, Afrah Alanazi, Inam Alanazi, Sarah Almaghrabi, Norah Alanazi and Samia Yahyaoui; data curation, Karim Gasmi, Sarah Almaghrabi, Norah Alanazi and Samia Yahyaoui; writing—original draft preparation, Karim Gasmi, Afrah Alanazi, Inam Alanazi and Sahar Almenwer; writing—review and editing, Sarah Almaghrabi, Inam Alanazi and Samia Yahyaoui; visualization, Karim Gasmi, Sahar Almenwer and Sarah Almaghrabi; supervision, Karim Gasmi and Inam Alanazi; project administration, Karim Gasmi, Afrah Alanazi, Norah Alanazi and Samia Yahyaoui; funding acquisition, Karim Gasmi. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this study are openly accessible at the following link: https://www.kaggle.com/datasets/anhoangvo/acdc-dataset/data.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. Cardiovascular diseases (CVDs)—Key facts. Key facts page indicates 19.8 million CVD deaths in 2022 (32% of all deaths). 2025 [cited 2026 Feb 8]. Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
2. World Heart Federation. World Heart Report 2023: confronting the world's number one killer. Reports 20.5 million CVD deaths in 2021 (nearly one third of global deaths). 2023 [cited 2026 Feb 8]. Available from: https://world-heart-federation.org/wp-content/uploads/World-Heart-Report-2023.pdf.
3. Petitjean C, Dacher JN. A review of segmentation methods in short axis cardiac MR images. Med Image Anal. 2011;15(2):169–84. doi:10.1016/j.media.2010.12.004.
4. Chen C, Qin C, Qiu H, Tarroni G, Duan J, Bai W, et al. Deep learning for cardiac image segmentation: a review. Front Cardiovasc Med. 2020;7:25. doi:10.3389/fcvm.2020.00025.
5. Iqbal T, Soliman O, Sultan S, Ullah I. Machine learning approaches for segmentation of cardiovascular neurocristopathy related images. IEEE Access. 2023;11:118301–17. doi:10.1109/access.2023.3325960.
6. Bernard O, Lalande A, Zotti C, Cervenansky F, Yang X, Heng PA, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans Med Imaging. 2018;37(11):2514–25. doi:10.1109/tmi.2018.2837502.
7. Isensee F, Jäger P, Full PM, Wolf I, Engelhardt S, Maier-Hein KH. Automatic cardiac disease assessment on cine-MRI via time-series segmentation and domain specific features. In: Statistical atlases and computational models of the heart (STACOM)—ACDC and MMWHS challenges. Berlin/Heidelberg, Germany: Springer; 2018. p. 120–9.
8. Kass M, Witkin A, Terzopoulos D. Snakes: active contour models. Int J Comput Vis. 1988;1(4):321–31. doi:10.1007/BF00133570.
9. Xu C, Prince JL. Snakes, shapes, and gradient vector flow. IEEE Trans Image Process. 1998;7(3):359–69. doi:10.1109/83.661186.
10. Yezzi A, Tsai A, Willsky A. A statistical approach to snakes for bimodal and trimodal imagery. In: Proceedings of the Seventh IEEE International Conference on Computer Vision; 1999 Sep 20–27; Kerkyra (Corfu), Greece.
11. Osher S, Sethian JA. Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. J Comput Phys. 1988;79(1):12–49. doi:10.1016/0021-9991(88)90002-2.
12. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models-their training and application. Comput Vis Image Underst. 1995;61(1):38–59. doi:10.1006/cviu.1995.1004.
13. Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. In: Computer Vision—ECCV'98. Berlin/Heidelberg, Germany: Springer; 1998. p. 484–98. doi:10.1007/bfb0054760.
14. Mitchell SC, Bosch JG, Lelieveldt BPF, van der Geest RJ, Reiber JHC, Sonka M. 3-D active appearance models: segmentation of cardiac MR and ultrasound images. IEEE Trans Med Imaging. 2002;21(9):1167–78. doi:10.1109/tmi.2002.804425.
15. Lorenzo-Valdés M, Sanchez-Ortiz G, Mohiaddin R, Rueckert D. Segmentation of 4D cardiac MR images using a probabilistic atlas and the EM algorithm. Med Image Anal. 2004;8(3):255–65. doi:10.1007/978-3-540-39899-8_55.
16. Lötjönen J, Kivistö S, Koikkalainen J, Smutek D, Lauerma K. Statistical shape model of atria, ventricles and epicardium from short- and long-axis MR images. Med Image Anal. 2004;8(3):371–86. doi:10.1007/978-3-540-39899-8_57.
17. Staib LH, Duncan JS. Boundary finding with parametrically deformable models. IEEE Trans Pattern Anal Mach Intell. 1992;14(11):1061–75. doi:10.1109/34.166621.
18. Chakraborty A, Staib LH, Duncan JS. Deformable boundary finding in medical images by integrating gradient and region information. IEEE Trans Med Imaging. 1996;15(6):859–70. doi:10.1109/42.544503.
19. Paragios N. A variational approach for the segmentation of the left ventricle in cardiac image analysis. Int J Comput Vis. 2002;50(3):345–62. doi:10.1023/A:1020882509893.

20. Paragios N, Rousson M, Ramesh V. Knowledge-based registration and segmentation of the left ventricle: a level set approach. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV); 2002 Dec 3–4; Orlando, FL, USA. p. 37–42.
21. Papademetris X, Sinusas AJ, Dione DP, Constable RT, Duncan JS. Estimation of 3-D left ventricular deformation from medical images using biomechanical models. IEEE Trans Med Imaging. 2002;21(7):786–800. doi:10.1109/tmi.2002.801163.
22. Pham QT, Vincent F, Clarysse P, Croisille P, Magnin IE. A FEM-based deformable model for the 3D segmentation and tracking of the heart in cardiac MRI. In: Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis (ISPA); 2001 Jun 19–21; Pula, Croatia. p. 250–4. doi:10.1109/ISPA.2001.938636..
23. Jolly MP, Xue H, Grady L, Gühring J. Combining registration and minimum surfaces for the segmentation of the left ventricle in cardiac cine MR images. In: Medical image computing and computer-assisted intervention (MICCAI). Vol. 5762. Cham, Switzerland: Springer; 2009. p. 910–8.
24. Idrissi Khaldi M, Erraissi A, Hain M, Banane M. Comparative analysis of supervised machine learning classification models. In: Intersection of artificial intelligence, data science, and cutting-edge technologies: from concepts to applications in smart environment. Cham, Switzerland: Springer Nature; 2025. p. 321–6. doi:10.1007/978-3-031-88304-0_44.
25. Carpenter S, Rawat R. Comparing machine learning classifiers for heart disease prediction: an empirical study of AdaBoost, KNN, and ANN with GLCM-based feature selection. Int J Progress Res Eng Manag Sci (IJPREMS). 2025;5(9):720–7.
26. Odeyemi J. Supervised learning showdown: kNN, SVM, neural networks, and boosted trees. Comparative analysis including biomedical datasets (e.g., fetal health); contrasts kNN, SVM, NN, AdaBoosted Decision Trees. 2025 [cited 2026 Jan 1]. Available from: https://jethroodeyemi.github.io/posts/2025/01/supervised-learning-analysis/.
27. Zhu C. Learning and application of different machine learning methods (KNN, SVM, Decision Tree) in different datasets. In: International Conference on Machine Learning and Computer Application (ICMLCA). Hangzhou, China: IEEE; 2024. p. 1–5. doi:10.1109/ICMLCA63499.2024.10754404.
28. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention (MICCAI). Vol. 9351. Cham, Switzerland: Springer; 2015. p. 234–41.
29. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Medical image computing and computer-assisted intervention (MICCAI). Vol. 9901. Cham, Switzerland: Springer; 2016. p. 424–32.
30. Milletari F, Navab N, Ahmadi SA. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV); 2016 Oct 25–28; Stanford, CA, USA. p. 565–71. doi:10.1109/3dv.2016.79.
31. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–11. doi:10.1038/s41592-020-01008-z.
32. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, et al. Automated cardiac MR image analysis with fully convolutional networks. J Cardiovasc Magn Reson. 2018;20(1):65. doi:10.1186/s12968-018-0471-x.
33. Qin C, Bai W, Schlemper J, Petersen SE, Piechnik SK, Neubauer S, et al. Joint learning of motion estimation and segmentation for cardiac MR image sequences. In: Medical image computing and computer-assisted intervention (MICCAI). Cham, Switzerland: Springer; 2018. p. 472–80.
34. Zheng Q, Delingette H, Duchateau N, Ayache N. 3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE Trans Med Imaging. 2018;37(9):2137–48. doi:10.1109/tmi.2018.2820742.
35. Khened M, Kollerathu VA, Krishnamurthi G. Fully convolutional multi-scale residual DenseNets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers. Med Image Anal. 2019;51(1):21–45. doi:10.1016/j.media.2018.10.004.
36. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, et al. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE Trans Med Imaging. 2018;37(2):384–95. doi:10.1109/tmi.2017.2743464.

37. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv:2102.04306. 2021. doi:10.48550/arXiv.2102.04306.
38. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-UNet: UNet-like pure transformer for medical image segmentation. In: Computer Vision-ECCV 2022 Workshops. Cham, Switzerland: Springer; 2023. p. 205–18. doi:10.1007/978-3-031-25066-8_9.
39. Zhao L, Zhou D, Jin X, Zhu W. Nn-TransUNet: an automatic deep learning pipeline for heart MRI segmentation. Life. 2022;12(10):1570. doi:10.3390/life12101570.
40. Mortada MJ, Tomassini S, Anbar H, Morettini M, Burattini L, Sbrollini A. Segmentation of anatomical structures of the left heart from echocardiographic images using deep learning. Diagnostics. 2023;13(10):1683. doi:10.3390/diagnostics13101683.
41. Leclerc S, Smistad E, Pedrosa J, Ostvik A, Cervenansky F, Espinosa F, et al. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Trans Med Imaging. 2019;38(9):2198–210. doi:10.1109/tmi.2019.2900516.
42. Liu F, Wang K, Liu D, Yang X, Tian J. Deep pyramid local attention neural network for cardiac structure segmentation in two-dimensional echocardiography. Med Image Anal. 2021;67:101873. doi:10.1016/j.media.2020.101873.
43. Moradi S, Oghli MG, Alizadehasl A, Shiri I, Oveisi N, Oveisi M, et al. MFP-Unet: a novel deep learning based approach for left ventricle segmentation in echocardiography. Phys Med. 2019;67:58–69. doi:10.1016/j.ejmp.2019.10.001.
44. Lei Y, Fu Y, Roper J, Higgins K, Bradley JD, Curran WJ, et al. Echocardiographic image multi-structure segmentation using Cardiac-SegNet. Med Phys. 2021;48(5):2426–37. doi:10.1002/mp.14818.
45. Merkow J, Marsden A, Kriegman D, Tu Z. Dense volume-to-volume vascular boundary detection. In: Medical image computing and computer-assisted intervention (MICCAI). Vol. 9901. Cham, Switzerland: Springer; 2016. p. 371–9.
46. Wolterink JM, van Hamersvelt RW, Viergever MA, Leiner T, Išgum I. Coronary artery centerline extraction in cardiac CT angiography using a CNN-based orientation classifier. Med Image Anal. 2019;51(4):46–60. doi:10.1016/j.media.2018.10.005.
47. Bortsova G, Bos D, Dubost F, Vernooij MW, Ikram MK, van Tulder G, et al. Automated segmentation and volume measurement of intracranial internal carotid artery calcification at noncontrast CT. Radiol Artif Intell. 2021;3(5):e200226. doi:10.1148/ryai.2021200226.
48. de Carvalho Macruz FB, Lu C, Strout J, Takigami A, Brooks R, Doyle S, et al. Quantification of the thoracic aorta and detection of aneurysm at CT: development and validation of a fully automatic methodology. Radiol Artif Intell. 2022;4(2):e210076. doi:10.1148/ryai.210076.
49. Berhane H, Scott M, Elbaz M, Jarvis K, McCarthy P, Carr J, et al. Fully automated 3D aortic segmentation of 4D flow MRI for hemodynamic analysis using deep learning. Magn Reson Med. 2020;84(4):2204–18. doi:10.1002/mrm.28257.
50. Pirruccello JP, Chaffin MD, Chou EL, Fleming SJ, Lin H, Nekoui M, et al. Deep learning enables genetic analysis of the human thoracic aorta. Nat Genet. 2022;54:40–51. doi:10.1101/2020.05.12.091934.
51. Marin-Castrillon DM, Lalande A, Leclerc S, Ambarki K, Morgant MC, Cochet A, et al. 4D segmentation of the thoracic aorta from 4D flow MRI using deep learning. Magn Reson Imaging. 2023;93(4):159–70. doi:10.1016/j.mri.2022.12.021.
52. Perrin S, Levilly S, Mouchère H, Serfaty JM. Super-resolution and segmentation of 4D Flow MRI using Deep learning and Weighted Mean Frequencies. In: International Conference on Medical Image Computing and Computer-Assisted. doi:10.1007/978-3-032-04965-0_52.
53. Mukisa R, Bansal AK. Cardiac MRI semantic segmentation for ventricles and myocardium using deep learning. In: Intelligent computing. Cham, Switzerland: Springer Nature; 2024. p. 169–88. doi:10.1007/978-3-031-62269-4_12.

54. Sun X, Cheng LH, Plein S, Garg P, van der Geest RJ. Deep learning based automated left ventricle segmentation and flow quantification in 4D flow cardiac MRI. J Cardiovasc Magn Reson. 2024;26(1):100003. doi:10.1016/j.jocmr.2023.100003.
55. Reiter G, Reiter C, Ovcina I, Fuchsjäger M, Reiter U. Four-dimensional flow MRI for a dynamic perspective on the heart and adjacent great vessels. Radiology. 2025;316(2):e242972. doi:10.1148/radiol.242972.
56. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: a nested U-Net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Cham, Switzerland: Springer International Publishing; 2018. p. 3–11. doi:10.1007/978-3-030-00889-5_1.
57. Lin TY, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 936–44. doi:10.1109/cvpr.2017.106.
58. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Computer Vision—ECCV 2018. Cham, Switzerland: Springer; 2018. p. 833–51. doi:10.1007/978-3-030-01234-2_49.
59. Ryu K, Lee C, Han Y, Pang S, Kim YH, Choi C, et al. Multi-planar 2.5D U-Net for image quality enhancement of dental cone-beam CT. PLoS One. 2023;18(5):e0285608. doi:10.1371/journal.pone.0285608.
60. Cai W, Wang B. DSE-net: deep semantic enhanced network for mobile tongue image segmentation. In: Neural information processing. Singapore: Springer Nature; 2023. p. 138–50. doi:10.1007/978-981-99-1648-1_12.
61. Liu H, Zhao B. Dynamic dual-stream feature fusion for semi-supervised medical image segmentation. In: Pattern recognition and computer vision. Singapore: Springer Nature; 2026. p. 360–74. doi:10.1007/978-981-95-5693-9_25.
62. Karthikeyan VD, Anusuya S. Comparative analysis of cardiac segmentation using custom U-Net on ACDC and cardiac catheterization datasets. In: Proceedings of the 2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA); 2025 Jun 25–27; Coimbatore, India. p. 1886–93. doi:10.1109/icirca65293.2025.11089794.
63. Mi Y, Zhang J, Jin H, Yin J, He Y, Xie G, et al. Dynamic thresholding and robust contrastive techniques for enhanced semi-supervised cardiac segmentation. PLoS One. 2026;21(4):e0342567. doi:10.1371/journal.pone.0342567.
64. Miao J, Chen C, Zhang K, Chuai J, Li Q, Heng PA. Cross prompting consistency with segment anything model for semi-supervised medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2024. Cham, Switzerland: Springer Nature; 2024. p. 167–77. doi:10.1007/978-3-031-72120-5_16.
65. Chaitanya K, Erdil E, Karani N, Konukoglu E. Local contrastive loss with pseudo-label based self-training for semi-supervised medical image segmentation. Med Image Anal. 2023;87(11):102792. doi:10.1016/j.media.2023.102792.
66. Yuan Y, Wang X, Yang X, Li R, Heng PA. Semi-supervised class imbalanced deep learning for cardiac MRI segmentation. In: Medical Image Computing and Computer Assisted Intervention—MICCAI 2023. Cham, Switzerland: Springer Nature; 2023. p. 459–69. doi:10.1007/978-3-031-43901-8_44.
67. Zhao Z, Wang Z, Wang L, Yu D, Yuan Y, Zhou L. Alternate diverse teaching for semi-supervised medical image segmentation. In: Computer Vision-ECCV 2024. Cham, Switzerland: Springer Nature; 2024. p. 227–43. doi:10.1007/978-3-031-72652-1_14.
68. Santos da Silva IF, Silva AC, de Paiva AC, Gattass M. A cascade approach for automatic segmentation of cardiac structures in short-axis cine-MR images using deep neural networks. Expert Syst Appl. 2022;197:116704. doi:10.1016/j.eswa.2022.116704.
69. Yang R, Yu J, Yin J, Liu K, Xu S. A dense R-CNN multi-target instance segmentation model and its application in medical image processing. IET Image Process. 2022;16(9):2495–505. doi:10.1049/ipr2.12503.